

BRIDGING THE GAP BETWEEN HEALTHCARE PROVIDERS AND  
CONSUMERS: EXTRACTING FEATURES FROM ONLINE FORUM TO MEET  
SOCIAL NEEDS OF PATIENTS USING NETWORK ANALYSIS AND  
EMBEDDINGS

Maitreyi Mokashi

Submitted to the faculty of the University Graduate School  
in partial fulfillment of the requirements  
for the degree  
Master of Science  
in the School of Informatics and Computing,  
Indiana University

August 2020

Accepted by the Graduate Faculty of Indiana University, in partial  
fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

---

Sunandan Chakraborty, Ph.D., Chair

---

Josette Jones, R.N., Ph.D.

---

Jiaping Zheng, Ph.D.

© 2020

Maitreyi Mokashi

## DEDICATION

I would like to dedicate my thesis research work to my father, who himself was a chronic disease patient and lost his battle to it. I could see and sense his struggle during his treatment and after but didn't quite understand it as I was too young. But today I understand him and his struggle a bit better. This work is my first step towards trying to make the lives of patients and survivors better through my skill and knowledge.

## ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my thesis advisor Dr. Sunandan Chakraborty for his invaluable guidance, support and introducing me to “The World of Data”. I was transitioning from a student who had just enrolled to get a Master’s degree to a student who started to envision a career in research. It’s because of his motivation, guidance and constructive criticisms that helped me mold into a research setting and generate a genuine liking for it. I truly appreciate his contribution of time, ideas and endless support.

I would like to thank my committee members Dr. Josette Jones and Prof. Jiaping Zheng for timely and invaluable advice. I also like to extend my heartfelt appreciation to Enming Zhang for her guidance and her inputs in my thesis research work.

I thank the School of Computing, IUPUI, the HCC department and Elizabeth Cassell for their support. I would also like to thank the professors of my different courses, my fellow classmates and friends. I learnt so much from each and every one of them in the past two years.

Lastly, I would like to thank my mother, Mona Mokashi and my aunt Dr. Meeta Pradhan, who are my pillar of strength. They are my role models and I inspire to be just like them: kind, thoughtful, professional and most importantly, an empowered woman!

Maitreyi Sameer Mokashi

BRIDGING THE GAP BETWEEN HEALTHCARE PROVIDERS AND  
CONSUMERS: EXTRACTING FEATURES FROM ONLINE FORUM TO MEET  
SOCIAL NEEDS OF PATIENTS USING NETWORK ANALYSIS AND  
EMBEDDINGS

Chronic disease patients have to face many issues during and after their treatment. A lot of these issues are either personal, professional, or social in nature. It may so happen that these issues are overlooked by the respective healthcare providers and become major obstacles in the patient's day-to-day life and their disease management.

We extract data from an online health platform that serves as a 'safe haven' to the patients and survivors to discuss help and coping issues. This thesis presents a novel approach that acts as the first step to include the social issues discussed by patients on online health forums which the healthcare providers need to consider in order to create holistic treatment plans.

There are numerous online forums where patients share their experiences and post questions about their treatments and their subsequent side effects. We collected data from an "Online Breast Cancer Forum". On this forum, users (patients) have created threads across many related topics and shared their experiences and questions. We connect the patients (users) with the topic in which they have posted by converting the data into a bipartite network and turn the network nodes into a high-dimensional feature space. From this feature space, we perform community detection on the node embeddings to unearth latent connections between patients and topics.

We claim that these latent connections, along with the existing ones, will help to create a new knowledge base that will eventually help the healthcare providers to understand and acknowledge the non-medical related issues to a treatment, and create more adaptive and personalized plans.

We performed both qualitative and quantitative analysis on the obtained embeddings to prove the superior quality of our approach and its potential to extract more information when compared to other models.

Sunandan Chakraborty, PhD, Chair

## TABLE OF CONTENTS

DEDICATION .....	iv
ACKNOWLEDGMENTS .....	v
ABSTRACT .....	vi
List of Tables .....	x
List of Figures .....	xi
List of Abbreviations .....	xii
CHAPTER 1 .....	1
INTRODUCTION .....	1
1.1 Motivation .....	1
1.2 Significance of work .....	3
1.3 Approach .....	4
1.4 Organization of the Thesis .....	6
CHAPTER 2 .....	8
BRIDGING THE GAP .....	8
2.1 Overview .....	8
2.2 Web-based tools for healthcare .....	8
2.3 Bridging the Gap: Case Studies .....	11
2.4 Summary .....	12
CHAPTER 3 .....	14
LITERATURE REVIEW .....	14
3.1 Social Media and Breast Cancer .....	14
3.2 Mining and Machine Learning on Network Data .....	16
3.3 Summary .....	17
CHAPTER 4 .....	19
DATA .....	19
4.1 Understanding the Data .....	19
4.2 Data Statistics .....	22
4.3 Previous work conducted on this data .....	23
CHAPTER 5 .....	25
PROBLEM DEFINITION .....	25



CHAPTER 6 .....	29
METHODOLOGY .....	29
6.1 Overview .....	29
6.2 Patient – Topic Model.....	29
6.3 Network Embedding .....	30
6.4 Community Detection .....	33
6.5 Summary .....	34
CHAPTER 7 .....	36
PATIENT – TOPIC MODEL EVALUATION .....	36
7.1 Overview.....	36
7.2 Experiment Setup.....	37
7.2.1 Hyperparameter Setup .....	40
7.3 Qualitative Analysis.....	41
7.3.1 Overview .....	41
7.3.2 Node Embedding Clustering using k-means.....	42
7.3.3 Community Detection: Observation and Inference .....	43
7.4 Quantitative Analysis.....	47
7.4.1 Overview .....	47
7.4.2 Baseline Models.....	48
7.4.3 Coherence .....	51
7.4.4 Comparison with reference dataset .....	52
7.4 Summary .....	54
CHAPTER 8 .....	55
CONCLUSION.....	55
8.1 Contributions .....	58
8.2 Future Work .....	60
Appendices.....	62
Publications.....	62
References.....	63
Curriculum Vitae	

## LIST OF TABLES

Table 4.1: Description of Online Breast Cancer Forum .....	22
Table 4.2: Analysis of post in each level of OBCF .....	23
Table 7.1: Count of users and topics after filtering .....	39
Table 7.2: Hyperparameters for the embedding model .....	41
Table 7.3: Topics in Cluster #01 .....	44
Table 7.4: Patients (users) in Cluster #01 .....	44
Table 7.5: Topics in Cluster #02 .....	46
Table 7.6: Patients (users) in Cluster #02 .....	46
Table 7.7: Comparing the performance of patient – topic network with baselines models using Normalized Pointwise Mutual Information (NPMI) to measure coherence.....	52
Table 7.8: Comparing the performance of patient – topic network with baselines models with respect to reference dataset in identifying communities.....	53

## LIST OF FIGURES

Figure 1.1: Gap between healthcare provider and consumer.....	4
Figure 1.2: Social Network Analysis: link connection .....	5
Figure 1.3: Latent feature space representation .....	5
Figure 4.1: Hierarchical structure of the data extracted from Breastcancer.org (OBCF) .....	20
Figure 5.1: Patient – Topic Network.....	26
Figure 5.2: Patient – Topic Feature Representation.....	27
Figure 6.1: DFS and BFS for a <i>node</i> where number of walks = 4, where start node is 1.....	32
Figure 7.1: ERD for BRCA database.....	38
Figure 7.2: Representation of the obtained communities from the embedding data.....	43
Figure 7.3: Network of users on online health platform.....	49
Figure 7.4: User model in representational vector space.....	49
Figure 7.5: User embedding mapped to topic space.....	50
Figure 7.6: Word2vec embedding using skip-gram.....	51

## LIST OF ABBREVIATIONS

1. BC: Breast Cancer
2. OBCF: Online Breast Cancer Forum

## CHAPTER 1

### INTRODUCTION

#### 1.1 Motivation

Chronic diseases such as heart disease, cancer, and diabetes are the leading causes of death and disability globally. According to the National Center for Chronic Disease Prevention and Health Promotion 6 out of 10 Americans live with at least one chronic disease (National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP), 2019). A diagnosis of a chronic disease means ongoing treatments, medications and lifestyle changes. The latter often also means physical, mental, and/or social adaptations. When rebooting and restarting their life as it were before is not an option anymore, it is difficult for the patient to cope with the new lifestyle and move forward. For example, women who are undergoing treatments for breast cancer find it difficult to find suitable life partners or even face break-ups and divorce as a result of the body image change. External factors such as these turn out to become major obstacles in their recovery, this sudden change in their way or lifestyle eventually deteriorates their health and specially their mental health. Hence, the patient's coping plays an equally important role as the treatments and medications.

Although the healthcare providers try their best to facilitate the optimal treatment plan for the patients it is majorly focused on the medical aspects of the treatment and the experience of the disease by the patient is often overlooked. Moreover, patients also refrain from discussing such issues due to the stigma associated with it (Nyblade et al, 2019). This lack of transparency between the healthcare provider and healthcare

consumer leads to further psychological deterioration in the respective consumers. It has become important for the healthcare providers to acknowledge and understand the personal, professional and social issues that the patient suffers as a consequence of the chronic disease. Major studies have yet to be conducted on how people integrate disease management, especially the ones suffering from chronic diseases, into their daily lives.

According to American Cancer Society Breast Cancer is the most common type of cancer in women next to skin cancer (American Cancer Society (ACS), 2020). They approximate that 1 in 8 women and 1 in 883 men are likely to be diagnosed with breast cancer in 2020. The overall death is decreasing by 1.3% per year (as per the statistics from 2013 to 2017) but, there are still more than 3.5 million breast cancer survivors in the USA. Hence, we conduct our study on the issues faced by Breast Cancer (BC) survivors. It is imperative for us to acknowledge, understand and work towards the betterment of the lives of these survivors.

Creating a holistic treatment plan is important for the patient's well-being and adaptation to disease management. The psychological stress that the patient suffers from as a result of the diagnosis and treatments should not become an obstacle in their well-being. How can we capitalize on ML features to shorten the gap between clinicians' understanding of a disease progress and a patient's experience of the disease? Which data can tell us the personal, professional and social problems faced by these patients in their honest and true form? That's where social media platforms plays an important role. The rapid growth of Web 2.0 has made social media a significant platform for health surveillance and social intelligence (Jin et al, 2016). The online health forum serves as a 'safe haven' to these patients where they seek solace and also, express themselves freely

without having any feeling of pressure, shame or bias. These platforms not only discuss the medical related issues like tests, medications, surgeries etc. but also discuss their personal and social issues. Using these platforms, patients form an interactive network by posting and replying to messages, providing reviews and attending discussion boards (Jin et al, 2016). Here they can post about their feelings of ‘not being okay’ and what they did to ‘feel okay’.

## 1.2 Significance

This thesis makes the following fundamental contribution to understand the existing gap between the providers and the patients based on the social issues related to the diagnosis and treatment. Connecting personal and social issues associated with a chronic disease, is important for disease management. There still exists a major gap between the healthcare providers and healthcare consumers when it comes to addressing the personal issues of the patients. The contextual factors of daily living with a chronic disease are not well understood; often overlooked in clinical care and hence, this leads to the decline in the patients’ health despite undergoing the treatment.

The online social media health forums serve as a platform to chronic illness survivors to share their experiences and thoughts related to their diagnosis and treatments. Through the data extracted from such online health platforms we are able to observe and understand the issues that these survivors have faced or are facing. This user – generated data creates an interaction network of the users and the issues they are discussing on the forum. This network will help us to extract the latent connections between the users and the context being discussed. *Latent* refers to *unknown* or *hidden*.

Unearthing these unknown connections will help patients with similar diagnosis to understand the foreseeable complications better. This will strengthen the healthcare cycle and help create a holistic healthcare ecosystem.

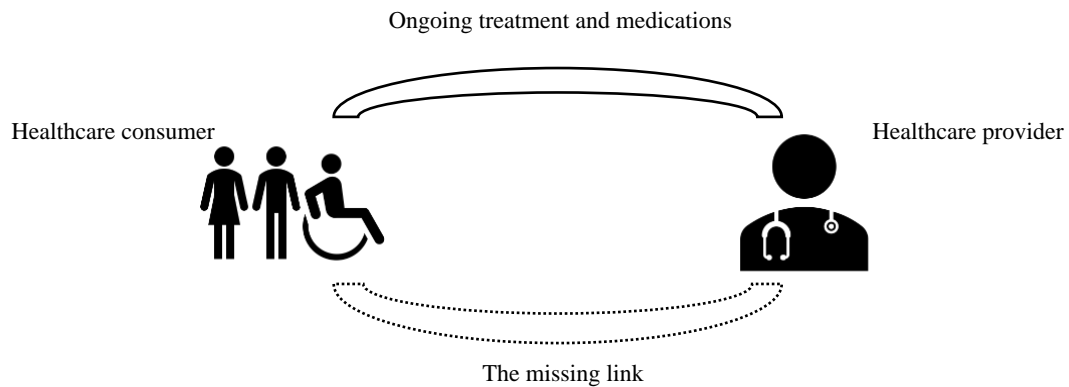


Figure 1.1: Gap between healthcare provider and consumer

### 1.3 Approach

In online social media platforms like Twitter, Instagram or Facebook, for example, if there is a post made by a user on “The Dwindling Numbers of Panda Population!”, which gains replies from two different users, say U-1 and U-2. U-2 has posted or replied to another post about the “Endangered Royal Bengal Tiger”. As the two users share some common interests, they have formed a connection through a hidden link. Due to their shared interests, we infer that U-1 is also interested in the population crisis of Royal Bengal Tiger along with the dwindling number of Pandas and we can infer that U-1 might get interested in this topic as well in the future (Figure 1.2).



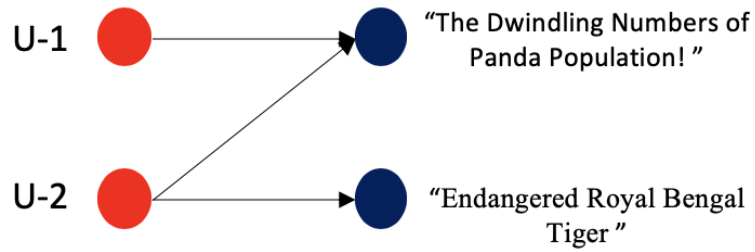


Figure 1.2: Social Network Analysis: link connection

Visualizing these two users and the topics they have interacted with, in a n-dimensional space, we observe that U-1 and U-2 are very close to each other along with their shared topic “The Dwindling Numbers of Panda Population!”. We will observe “Endangered Royal Bengal Tiger” closer to U-2 and close to U-1 hinting at a hidden association between the two (Figure 1.3). This is known as *latent feature representation*, where, latent refers to the *hidden relations* and the n-dimensional space is known as *feature space*.



Figure 1.3: Latent feature space representation

We applied the same concept to social media health forums. The interactions these users have with each other give us a direct idea about their association but there still is a lot to understand and learn from the indirect associations which need to be unearthed.

In some cases, this additional information might even be life-altering or lifesaving. Social media health forums have a similar basic organization structure as their entertainment counterparts. Patients (users) respond via posts in the *topics* that they are affiliated to and hence, form a hidden connection with the other patients in that topic thread. This connection when visualized in the *feature space* will help us understand the close proximity between certain patients and topics. These topics may not be of current interest to the patients but due to the proximity with *similar* patients, may be of importance in the future. This way, we can forewarn the patients of the various issues they might face in future and hence they can be properly equipped to face it. This is the *latent feature space representation* for patients in an online health forum.

## 1.4 Organization of the Thesis

**Chapter 2** reviews few of the existing tools/platforms/apps that have been created to try and bridge the gap between the healthcare providers and consumers as well as various studies conducted to address this issue. In this chapter we discuss the ideas and principles behind the existing platforms and studies. We elaborate on these principles and summarize by explaining how our idea, approach and implementation is different.

**Chapter 3** provides a brief description of the literature review conducted on published papers who have worked on social media and network analysis, text mining and healthcare data as well as breast cancer and machine learning. **Chapter 4** describes the data that we have extracted to implement our model. **Chapter 5** states the problem definition. In **Chapter 6**, we discuss the methodology used to conduct this study. In this chapter a detailed description of the patient – topic network, embedding techniques as

well as downstream tasks for model analysis have been presented. **Chapter 7** discusses the experiment setup, qualitative analysis as well as the quantitative analysis observation, inference and results. This chapter also discusses the other experiments that were conducted and used baselines models for evaluating our patient – topic bipartite network model. Finally, this thesis is concluded in **Chapter 8** with insights into the contributions of this work to the field as well as discusses the future scope of this research study.

## CHAPTER 2

### BRIDGING THE GAP

#### 2.1 Overview

Studies have been conducted to address and analyze the existing issue of “Gap between the healthcare providers and healthcare consumers”. Although these studies address different aspects of the *gap*, it is essential to understand them and determine what issue they address as well as their significance. In this section we will go over few of the existing healthcare platforms that have tried to shorten this gap as well as published works that have been conducted to get insights into the provider – patient relation.

#### 2.2 Web-based tools for healthcare

Web-based tools play a key role in bridging the gap between the healthcare providers and consumers. The modern day “internet world” that helps us connect with friends and family all over the world is now serving a purpose of connecting providers with patients for a smooth and efficient course of action. These web-based platforms all follow their respective ideology and they share a common thought of bringing the patients EMR data at one place.

##### 2.2.1 The Ascom Myco 3 by ascom for Healthcare

It is important between physicians, care teams as well as the patients to be on the same page. Seamless flow of clinical information between the respective personnel’s and devices is critical. It is often observed that different pieces of a patient’s medical history get stored in different locations and when most needed, most of the information is

inaccessible to the healthcare provider sometimes causing fatalities (George Palma, 2020 “*Electronic Health Records: The Good, the Bad and the Ugly*”)

A device that can help overcome this challenge is the Ascom Myco 3 smartphone (*Bridging clinical information and communication gaps for better-informed care*, 2020), which is purpose-built for healthcare and provides clinicians and caregivers with instant access to all applicable patient information. This device is a part of the Ascom Healthcare platform which provides modular healthcare information and communication solutions that integrate with existing information systems. By bringing together clinical care data often managed over multiple devices and systems, Ascom Myco solutions make patient care needs, notifications and updates more visible—and keeps them close to hand for clinicians. Ascom Myco 3 solutions connect clinicians and teams to enable care management, communication and coordination from the point of care to anywhere.

### **2.2.2 Veta Health**

The Veta Health Platform (Thakkar, 2008) emphasizes on the trust factor between physicians and patients. “*Veta Health is the agent of change at the forefront of understanding how patient and provider communication can be improved using technology and is focused on producing better health outcomes.*”

This platform focuses on several studies conducted to understand the effects of proper communication between healthcare providers and receivers. Effective patient provider communication can improve a patient’s health as much as many drugs can (Stewart, 1995). Their study is also based on another randomized controlled trial study (Bull et al, 2002) on patient-provider communications. It reported that the

quality of communication in both the history-taking and management-discussing portions of interactions resulted in patient outcomes that were influenced. The outcomes most impacted were emotional health, symptom resolution, and pain control. Most of the study resulted in a positive correlation between effective communication and improved patient health outcomes.

Veta Health is a platform with automated and responsive care pathways that are triggered based on patient input. Includes tools for self-management through educational resources, remote monitoring and real-time actions pushed to patients. Their robust data infrastructure captures and analyzes clinical and patient-generated data to drive appropriate treatment protocols and includes data from EMRs, consumer wearables, questionnaires, and patient-reported outcomes thus, have a relation with their providers outside the clinical environment. This platform also discusses the advantages of providing a platform where the patients feel involved in their disease management. For this purpose, they conducted a study on a group of patients who were identified as having congestive heart failure and observed the readmission rate of patients who have enrolled in this platform as opposed to those who haven't. Their study showed that the readmission rate of patients enrolled in this platform decreased by 75% as opposed to the control group.

### **2.2.3 iVEDiX**

(Ivedix, Mobile Healthcare Bridges Providers-Patient Gap, 2016) Some of the transformation of healthcare has focused on the adoption of patient-facing healthcare applications and providing patients with easy access to their healthcare data. Healthcare

providers would like to be able to access patient data from one mobile device quickly. Often, healthcare providers must use multiple systems to find all of the information they need about the patient. Physicians, hospitals, and other healthcare providers can use mobile technology to gain better access to patient medical records, records that include patient medical history, current prescriptions, hospital visits, etc. The iVEDiX physician rounding application securely accesses information from multiple systems and EMRs allowing clinicians to obtain a unified view of patients. The ability to quickly access and analyze patient data allows doctors to serve their patients better helping to improve their overall health and quality of life.

### **2.3 Bridging the Gap: Case Studies**

Authors Francis et al (1969) conducted a study of 800 outpatients visits to Children's Hospital of Los Angeles to explore the interaction between doctor and patient on patient satisfaction. This study showed that a lack of warmth in the interaction between the patients and their healthcare providers as well as failure to receive a proper explanation of their diagnosis and treatment from them became a major factor in the noncompliance from the patients end. Their follow-up survey also depicted this dissatisfaction numerically as 24% of the patients were grossly dissatisfied, 38% moderately satisfied and 11% noncompliant.

In lower-to-middle income countries, where there is already a lacking healthcare infrastructure this gap exists predominantly when it comes to understanding the patients socio-economic background for creating a treatment plan. They need to take into account the patient's difficulties to access care and obtain the proper resources for better disease

management. Goudge et al (2009) conducted a household survey of approximately 30 households over 10 months with descriptive narratives that helped gain textual data to understand the interactions with the health system. Of the cases, 34 cases were chronic illness, only 21 (62%) cases had an allopathic diagnosis and only 12 (35%) were receiving regular treatment. Livelihoods exhausted from previous illness and death, low income, and limited social networks, prevented consultation with monthly expenditure for repeated consultations as high as 60% of income.

- Interrupted drug supplies, insufficient clinical services at the clinic level necessitating referral, and a lack of ambulances further hampered access to care.
- Poor provider-patient interaction led to inadequate understanding of illness, inappropriate treatment action and the patients eventually giving up on the health system.

However, productive patient-provider interactions not only facilitated appropriate treatment action but enabled patients to justify their need for financial assistance to family and neighbors, and so access care. In addition, patients and their families with understanding of a disease became a community resource drawn on to assist others.

Authors Mira et al (2012) also conducted a similar study in the 14 health centers belonging to 3 primary care districts and 3 hospitals in Spain. Their study included 764 patients and 327 physicians to determine whether patients consider the information obtained from the physicians enough. Their study too showed that patients are not normally informed about medication interactions, precautions and foreseeable complications. The information provided by general practitioners does not seem to contribute enough to the patient involvement in clinical safety.



Often patients with long-term health issues develop multiple conditions and hence have to follow different treatments simultaneously. In the study conducted by Neuner-Jehle et al (2017) to investigate how well the general practitioners in Switzerland perceived the main complaints of their patients with several diseases. For this purpose, they compared 128 main complaints listed by patients to those listed by their general practitioners and investigated what factors influenced the degree of agreement or disagreement. Thus, a majority of general practitioners perceive the main complaint of the long-term patient correctly, but there is room for improvement as 21% of general practitioners do not list the chief complaints of the patient at all in a four-part list. To improve awareness of the multimorbid patient's most demanding illness, the authors recommend the practitioners to ask for the patient's chief complaints during every encounter with the patient and to provide room (including time) for dealing with the answer. From their study they also found that the most common complaint was pain and how to manage pain should be considered in a multifaceted approach.

It is a well-known fact the medical terms used by the healthcare providers during their interactions with the patients is at times not at par with the patient's level of grasping medical terminologies. This leads to the patients to not understand their diagnosis and treatment plan well. Authors Gu et al (2019) in their study addresses the issue of vocabulary gap between consumers and professionals in the medical domain that hinders information seeking and communication between the two parties. Their main objective was to create a machine learning solution to develop a method to identify and add new terms used by consumers to health vocabularies, which help the healthcare providers explain the treatments to their patients in a way that is comprehensive to them.

They created word embeddings from the text generated by users while browsing using an unsupervised method as well as created embeddings from clinical texts using supervised learning methods. The authors then created word pairs from these two embeddings which have similar meanings. With their method they could correctly identify 80% of the synonyms by just searching the top 10 similar words. This study was conducted for both English and Chinese language to give importance to locally used languages.

## **2.4 Summary**

In this chapter we discussed few of the existing platforms that are trying to bridge the gap between the healthcare providers and consumers by trying to collect all of the patient's medical documentations like past hospitalizations, diagnosis, treatments as well as capture the real-time data of the patients through the consumer wearables at one place, so as to not create confusion and commotion at the last minute. By surveying these online platforms, we observe that there is a need to develop a platform that provides additional information regarding the social issues faced by the patients with similar diagnosis.

The published works have conducted their respective studies to take a survey of the patient satisfaction level for their healthcare providers. These studies deduced the need for considering the patients social, personal or professional issues to be important for disease management depending on factors such as socio-economic background of the patient, the level of the respective patients medical/his diagnosis understanding as well as providing a safe space for the patients to open up to the providers and share their feelings without hesitation among others.

There is a need for better understanding of daily living with a chronic disease which is crucial for disease management. Yet current clinical practice focuses on signs and symptoms of a disease not on what it means to live with the disease. Collection of data on disease experience is not well established in clinical practice. Social media where patients and caregivers express their experiences freely may be a solution. Yet extracting knowledge from the postings is not easy. Our study is an attempt to create a machine learning solution to observe and understand the issues discussed by patients/survivors on health forums and analyze this user-generated content to create connections between the various contexts being discussed.

## CHAPTER 3

### LITERATURE REVIEW

#### 3.1 Social Media and Breast Cancer

During the past couple of years, many health social media websites have been created to facilitate information exchange among patients as these online forums and social media platforms allow patients to search for health information online and interact with people with similar conditions. Some websites like Everyday Health (<https://www.everydayhealth.com>) and PatientsLikeMe (<https://www.patientslikeme.com>) cover general health problems and provide information on many aspects of health. Others focus on a particular group of people with similar conditions, such as diatribe's focus on diabetic patients or Disabilities-R-Us' (<https://www.disabilities-r-us.com>) focus on people with disabilities. Many past researches have been conducted to analyze the effect of these online activities and how they can provide patient care and welfare. There are studies that have been conducted which advance our understanding of how social media helps patients with advice, guidance, and support with their chronic diseases. Greene et al (2011) used qualitative methods to evaluate the content of Facebook groups dedicated to diabetes management. This study concluded that a "safe" place to discuss extra-clinical issues helped the patients in general.

Apart from these qualitative studies, other observational studies have used data mining and machine learning in their analysis. For example, Park and Ryu (2014)

applied Natural Language Processing (NLP) methods to extensive online forum text data to understand key problem areas of patients who have fibromyalgia. A similar approach has been used to address a variety of clinical problems, such as public sentiments towards vaccination (Joshi et al, 2018), adverse drug effects (Jiang and Zheng, 2013), influenza epidemic using Twitter data (Aramaki et al, 2011), and e-cigarette usage (Zhan et al, 2017).

Chawla et al (2013), emphasized on how "*Data-driven and networks driven thinking and methods can play a critical role in the emergence of personalized healthcare.*" They use a patient-centric model (CARE) that creates a personalized disease risk profile, as well as a disease management and wellness plan for an individual.

Machine learning and text mining has been extensively used for breast cancer research.

Bodicoat et al (2020) and Grant (2020) conducted studies that use clinical data to better diagnosis and treatment of breast cancer. Many studies have been conducted for early detection of BC which include Dhahri et al (2019), Kourou et al (2015), Ragab et al (2019) and Shen et al (2019). Goldstein et al (2020) as well as Wang et al (2020) focused their research on various medical factors such as drugs and treatments for BC. In parallel, numerous studies have focused on the effects of social media on breast cancer research and patients. Modave et al (2019) performed sentiment analysis on tweets discussing breast cancer and demonstrated that social media can improve the perceptions of the disease in the general population. Zhang et al (2017) used CNNs to extract longitudinal information to understand the key topics discussed on online breast cancer forums. There are many factors that determine the proper survival of breast cancer patients, this included access to treatment, financial constraints and many more personal factors. The

study conducted by Sheng et al. (2019) focused on the long-term effects of these factors on the quality of life of breast cancer patients and survivors.

Trans-disciplinary research has been conducted to create frameworks that take into account social determinants of cancer to observe the environmental, social and behavioral factors of the cancer patients. Hiatt et al (2008) designed a framework to conceptualize how social determinants interact with other factors in the etiology of cancer and to capture changes over time. Carter et al (2009) conducted a qualitative analysis of thirty-two cancer control policy documents, critiquing them based on their likely impact on social determinants and created a matrix and set of questions to guide the development and assessment of health policy. Asch et al (2015) in their article state the importance of systematic surveillance of patients' social media data as it has the potential for the clinicians and the hospitals to obtain information that they have been missing or overlooking to date, which is a window into the day-to-day lives of their patients. They also state that mining social media data is the new frontier for precision and personalized medicines.

### **3.2 Mining and Machine Learning on Network Data**

Data extracted from health forums as well as other user-generated content are noisy, inconsistent and this makes it difficult for information extraction. As a result, several data mining techniques have been devised for such data mining tasks. Yang et al (2012) compare their study with Rossetti et al (2011) , where the later proposed multidimensional versions of the Common Neighbors and Adamic/Adar, and derived predictors that aimed at capturing the multidimensional and edge level temporal

information, while the prior gathered nodal historical data to capture the preference of topological features when two nodes are associated by new link; while they are interested in edge level communication data.

Network analysis plays an important role when it comes to understanding the connection formed between same or different entities who share similar interests for portray similar traits. It is difficult to make these connections from complex networks such as the ones extracted from online social media platforms. The similar interests shared between these entities are depicted using a link which serves the same purpose a rope plays while connecting the two poles. These links also help to unearth the hidden or unknown connections from the respective networks. Grover et al (2016) in their paper with the help of these links enabled the extraction of unknown or hidden relations from such networks using the Skipgram embedding technique of Mikolov et al (2013) as well as the random walk technique of DeepWalk by Perozzi et al. (2014).

Hu et al (2019), Kim et al (2018), Li et al (2017) and Peng et al (2019) conducted their research by building over the original node2vec model to meet the requirements of the respective studies. For example, Li et al (2017) created a modified version of node2vec by introducing TDL2vec which considers time factor while generating the links during word2vec model. Similarly, Peng et al (2019) created a model to predict Parkinson's disease by creating N2A-SVM algorithm which includes an autoencoder for dimensionality reduction of the node2vec model and Support Vector Machine for the prediction analysis. We propose to use the Node2Vec model to create the desired node embeddings on the bipartite network.

### 3.3 Summary

In this section we have studied the different research works that have been conducted in the field of social network analysis, natural language processing as well as various works conducted on breast cancer research using machine learning. Most of these works related to breast cancer include:

- Better diagnosis and early detection of breast cancer
- Understanding various factors like drugs, treatments as well as financial constraints on the patients.
- Sentiment analysis on tweets discussing breast cancer etc.

The published works in this section especially the ones address healthcare problems, don't stress the importance of creating approaches that intend to solve these issues taking into consideration the structure of the user-generated data and the need to inculcate the heterogeneous nature of the networks to try and solve the existing issues. All these studies and machine learning techniques have laid a foundation for our thesis, where we aim to find the social issues faced by breast cancer patients by extracting data from an online breast cancer forum and unearth the latent features to understand the social issues related to BC. Our method is designed taking into account the heterogeneous nature of the user - context network from an online health forum to get deeper insights into the social issues related to the respective chronic disease which is BC for our case study.



## CHAPTER 4

### DATA

#### 4.1 Understanding the Data

BC patients undergo many ordeals during and after their treatment. As mentioned in *section 1.1*, it is predicted that 1 in 8 women will be diagnosed with this chronic disease and that there are over 3.5 million current patients and survivors in the United States alone. It is important that the questions that arise in the minds of these patients are acknowledged. A lot of these questions are medically related, for example, “is this clinical trial working?”, “is it okay to undergo the recommended surgery?” or even “what are the side-effects of this medication?” but a lot of these questions are related to the patients/survivors day-to-day life like “Will I find a suitable partner?”, “He wants a divorce? How do I move on?”, “The medications are too expensive. How do I manage with my income?” Hence, it is important to extract data from an online platform which discusses both medical and non-medical issues related to BC.

In a previous work conducted by Jones et al (2018), they conducted QCA of 5 online breast cancer forums to get a better understanding of the context of the discussions on these forums. They had specific requirements that needed to be fulfilled before they approached the forum organizers with the data extraction request. An ideal forum should have at least 5000 registered users and at least 50,000 posts during the week of data crawling. It was important that the posts on this forum be sorted into various categories and not be in a general question – answer (QnA) format.

Breastcancer.org was chosen from the list of shortlisted websites as it had the maximum number of posts and met all the requirements. Breastcancer.org community provides a platform for the patients and their friends and family to share their experience and post questions etc. This platform hosts multiple forums which are generally specific to one subject (Jones et al, 2018) and users create new threads to post their questions or opinions related to a specific topic.

The levels of this online health platform (Figure 4.1) is explained below. The dataset consists of four levels excluding the users.

- **Level I - Categories:** The OBCF identifies each post as one of 9 different sections for surface-level categorization. This gives the users an overview of the discussions being held in the subsequent forums and topics.

- Day-to-Day Matters
- Not Diagnosed but concerned
- Advocacy and Fund-Raising
- Community Connections
- Welcome to Breastcancer.org
- Site News and Announcements
- Connecting with Others Who Have a Similar Diagnosis
- Moving On & Finding Inspiration After Breast Cancer

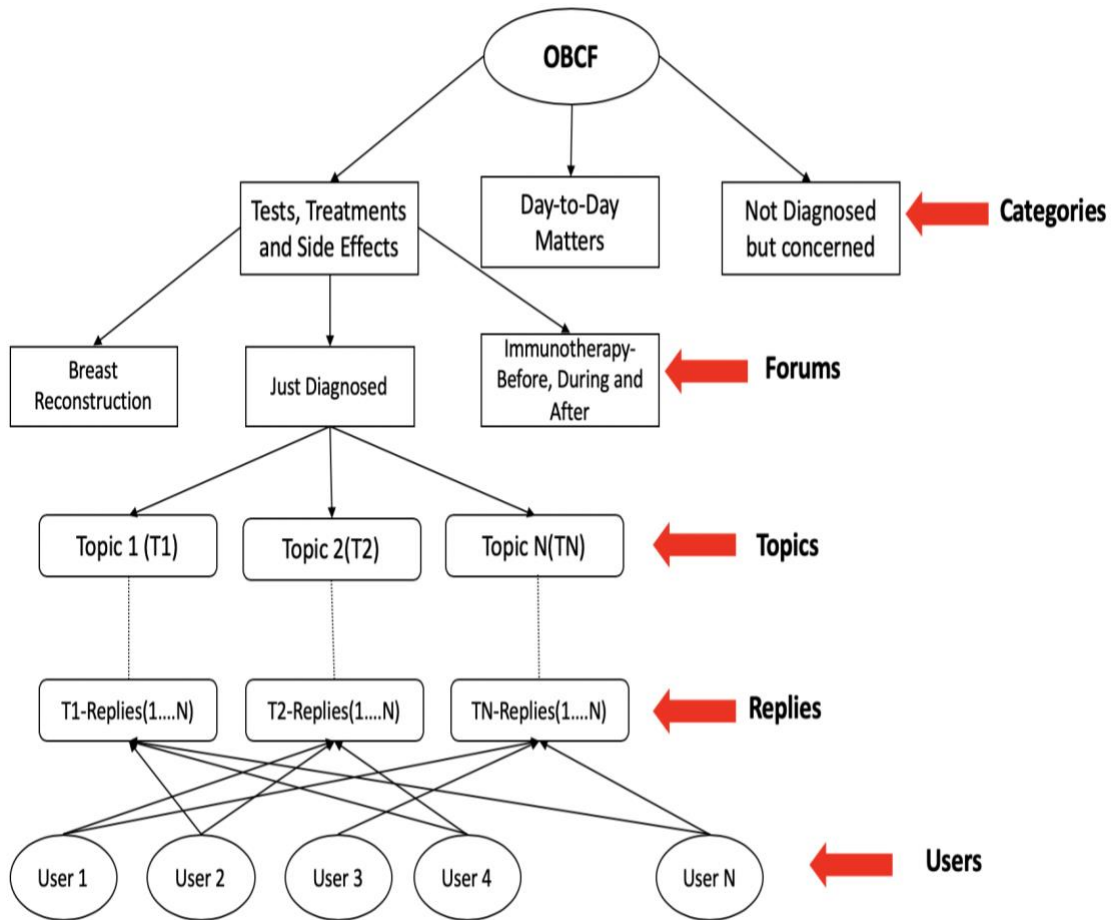


Figure 4.1: hierarchical structure of the data extracted from Breastcancer.org (OBCF)

• **Level II - Forums:** As each ‘Category’ gives us an overview of Breastcancer.org, ‘Forums’ further sorts them into selective discussions. For example, discussion regarding ‘mastectomy’ and ‘lumpectomy’ will most probably be addressed in the ‘Breast Reconstruction’ forum or problems such as ‘dating’ and ‘break-ups’ will be discussed in the ‘Day-to-Day Matters’ forum. We analyzed data from 79 out of the 81 forums, as the other 2 forums had corrupted user-IDs’. Out of approximately 4.4 million posts, one forum "Day-to-Day

Matters" has 616,598 which is the maximum number of posts in one forum. The average number of posts per forum is 56,300 posts. From this we observe a range of different issues the patients/survivors have to face on a daily basis and the importance addressing them. On the other hand, this also indicates that patients see this platform as a safe place to share their experiences with others.

- **Level III - Topics:** Any post that is not a reply becomes a new topic. When a user has a new subject to discuss or something new to share, they post an independent post which directly converts as a new topic. In Breastcancer.org, there are 140,000 topics spread across the 79 forums with a maximum of 56,000 replies for one individual topic and a mean of 30 replies. There are a few topics with no replies and have been excluded during data extraction for the final model.

A topic can be:

- "chemo after a mastectomy"
- "Size of tumor by MRI vs Reality" etc.

- **Level IV - Replies:** This level in the hierarchy consists of subsequent replies to individual topics. One user has made over 48,000 posts and the mean number of posts per user is 47. There is no sub-branching for the replies and all the replies to a particular topic are stacked (with a respective ID).

## 4.2 Data Statistics

Table 4.1: Description of Online Breast Cancer Forum

	<b>Total</b>
Categories	9
Forums	79
Topics	140,000
Replies	~4.4 million
Users	~94,000

Table 4.2: Analysis of post in each level of OBCF

<b>Posts per type</b>	<b>Forums</b>	<b>Topics</b>	<b>Users</b>
Max	616598	56091	48986
Min	11	1	1
Mean	56304	31	47

## 4.3 Previous work conducted on this data

Jones et al. (2018) extracted data from Breastcancer.org to determine the feasibility of acquiring and modeling the topics of this online breast cancer forum. Using qualitative content analysis or QCA they surveyed 5 different online breast cancer forums and concluded Breastcancer.org the best one to conduct their case study. They obtained

topic models and the obtained topics were placed into 4 distinct clusters using NLP and statistical modelling. The final topic model organized >4 million postings into 30 manageable topics. Finally, these 30 topics were grouped into 4 distinct clusters with similarity scores of  $\geq 0.80$ ; these clusters were labeled Symptoms & Diagnosis, Treatment, Financial, and Family & Friends. A clinician review confirmed the clinical significance of the topic clusters. They also performed a Machine Learning – Regression algorithm based on Akaike information criterion to identify the most significant topics across individual forums and demonstrated that 6 topics ranging from  $-642.75$  to  $-412.32$ —were statistically significant. The obtained result was clinically asserted as significant. The topic modeling was performed using Machine Learning Language Toolkit (MALLET) (McCallum, Andrew Kachites, 2002) an open source software, followed by multiple linear regression (MLR) analysis to detect highly correlated topics among the different website forums.

Zhang et al. (2019)\* manually annotated 736 randomly selected posts from Breastcancer.org and created “Patient-centered Thesaurus of Chronic Survival (PACToCS)”, which was then mapped with medical controlled vocabulary - NCI Metathesaurus. The authors identified 30 topics and 27 out of 323 full code terms from PACToCS matched with the full term of the NCI - Metathesaurus. They obtained a precision of 85% upon classification by multiple ML models.

We aim to create our model by preserving the original structure of the Breastcancer.org platform. Preserving this original structure will help us create our problem definition and methodology to unearth new findings and understand the explicit

relation i.e. relation between patient and topic as well as implicit relations that is the relation between similar nodes i.e. patient-patient and topic-topic nodes.

## CHAPTER 5

### PROBLEM DEFINITION

This thesis aims to extract features to represent patients from an online health forum based on their postings. These extracted features will give us an overview of the issues related to a disease state or diagnosis based on the descriptions provided by the patient (user) postings. We unearth these features from the information shared by them using their interaction pattern in the OBCF.

We assume that a patient has a direct relationship with a topic in which they have participated, where participation is defined as either posting a new message or replying to an existing one. We formulate the problem with three variables:

- Patients ( $V$ )
- Topics ( $T$ )
- Features ( $\theta$ ).

Here,  $V$  and  $T$  can be observed from data, whereas  $\theta$  is latent and will be learned.

Our objective is to design a mapping function that encodes patients, their messages and their interactions into a common feature space.

Our technique aims at binding patients and the related textual content together, i.e. similar patients with similar interests will be placed in close proximity in the high dimension feature space. This similarity is defined by the interaction of the patients in the forums – similar patients will tend to participate in the same or related topics.

We convert the forum data into a bipartite graph (Figure 5.1). In this graph, we represent the data as network  $G$ :

$$G = (V, T, E)$$



where,  $V$  represents the set of patients and  $T$  the set of topics extracted from the forum data.

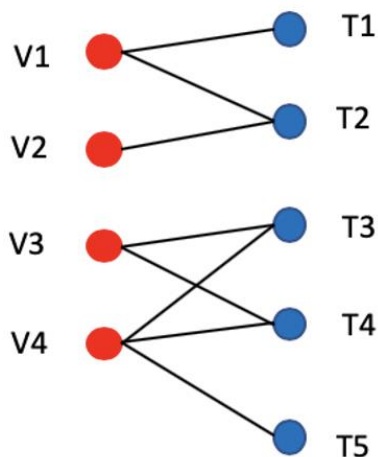


Figure 5.1: Patient-Topic Network

A patient  $v_i$  and a topic  $t_j$  will be connected by an edge ( $E$ ):

$$e_{ij} \in |V| \times |T|$$

$G$  is a bi-partite network hence, there will be edges between patients and topics, but no edge will exist between topic or patient pairs. The relationship between two patients or topics is implicit and identifying those implicit relationships is our main goal.

This technique will enable us to learn the features through the interaction between patients and topics and also using patient-patient relationships based on participation in common topics. For example, from Figure 5.1 User V1 and User V2 have posted in a common topic  $T2$  and User V1 has posted in another topic  $T1$ , is it possible for User V2 to be in the same community and have a connection with  $T1$ ? Is it possible for two or more topics whose contextual nature is different, but can they be included in the same due to the embedding technique? Can we unearth the hidden connection between such

nodes and understand a pattern in their commonalities that will help the users in their road to feeling better?

Our goal is to design a mapping function  $\psi$ , such that

$$\psi: V \cup T \rightarrow R^K$$

for all patients  $i$  and topics  $j$ . Here,  $K$  is a feature space of predetermined dimension.

We use an adaptive node embedding method to formulate this mapping. The mapping function  $\psi$  will produce  $K$ -dimensional vectors for each patient and topic with an embedding for the two respective variables which will then be represented in a  $n$ -dimensional feature space as shown in Figure 5.2.

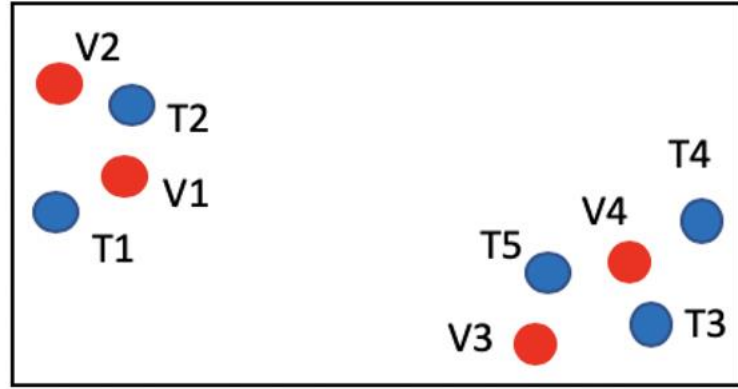


Figure 5.2: Patient-Topic Feature Representation,  
where V are the users and T are the topics

This way we will compute the distance between two patients, two topics as well as the distance between a patient and a topic. As a result, we are able to discover implicit as well as the explicit relationships within the network variables that were not directly depicted by the underlying network.

From the obtained clusters, we can analyze the commonalities between patients in a community and try to understand the issues that are observed in some of them and connect them to the other patients who have not been diagnosed with the said issues yet or make them aware about the same. At the same time, we will be able to directly connect patients with nearby topics and discover potential issues a patient might face given the current state of the respective patient.

Our novel approach of extracting the latent features by converting the data into a bipartite network to *incorporate the patients as intermediate variables* into the model has an advantage of obtaining deeper understanding of the user – generated network. Information shared by a single patient about a specific topic  $t$  is unlikely to contain all possible information about that topic. However, aggregating different patients' experience about  $t$  from the entire data can provide a more holistic overview of  $t$ .

Thus, including patients as intermediate connecting points is more likely to link many latent connections between the topics and the patients. Our fundamental contribution is that this network-driven method will lead to richer feature extraction compared to a purely text-based approach or a homogeneous model consisting of only topic or user nodes that will not include intermediate variables for deeper insights.

## CHAPTER 6

### METHODOLOGY

#### 6.1 Overview

We aim to represent breast cancer patients' state using latent features by analyzing the user interactions on an online forum. We represent the forum data as a bipartite network that represents the relationship between patients and topics shared on those forums. This network gives us an indirect relationship amongst the patients and those between the topics themselves respectively, that are not directly depicted in the network. Through our method we wish to quantify the patients' experience across different stages of the disease and encode that information into a high-dimensional vector that can embed a variety of information, including diagnosis, treatment, side effects of the treatment as well as mental/social issues on a personal level. This representational model will capture the latent features of the patients showing similar traits in a way that will help the medical practitioners plan improves and better personalized treatment plans for their patients without having the necessity to have the patient's complete identifiable biographical data.

#### 6.2 The Patient – Topic Network

Every breast cancer patient experiences a unique set of challenges. To connect the different issues related to a certain diagnosis or disease state from the user postings, we need to create a network that will connect these two entities. For example, a patient  $X$  shares their experience related to a treatment option  $Z$ , when the disease state is at  $Y$ ; we

aim to identify all issues related to Z across the dataset that the patients have mentioned in their postings.

In our dataset, the information related to a single post about Z and Y are represented in one topic, say  $t$ . In addition, our goal is to capture other issues shared by patients who are similar to X. This will provide a more **holistic view** of the issues faced by a patient who is in the same state as Y and who are receiving the treatment Z.

The desired latent features are learnt from the patient's interactions on the online forum. Thus, we convert the data into a patient-topic bipartite network ' $G$ ' and call this the patient-topic network. We represent the forum topics as  $T$ . A topic  $t_j \in T$  represents a new post and its subsequent stack of replies. ' $T$ ' or the topics are not automatically extracted from forum text using any machine learning algorithm but are the topics as defined within OBCF (section 4.1). A topic  $t_j$  will have  $T_j^N$  posts, including the original post and  $T_j^{N-1}$  replies.

$$V = \{v_1, v_2, v_3, v_4, \dots, v_M\},$$

represents the set of  $M$  patients, thus, there are  $M + N$  nodes in the patient-topic network  $G$ .

If a patient  $v$  has posted in  $t_j$ , i.e. has posted

$$t_j^k \in t_j,$$

where  $1 \leq k \leq t_j^N$ , there will be an edge between  $v$  and  $t_j$ .

### 6.3 Network Embedding

The desired vector representation will place similar nodes i.e. patients and topics closer to each other while placing others far apart in the high dimension vector space. As

our network consists of two types of nodes namely, patient (user) and topics, we need to construct a heterogeneous network to preserve the nature of the network. Unlike homogeneous networks, in a bipartite network the same type of nodes are not connected by an edge. In our case, two topic nodes or two patient nodes do not have an edge between them. However, that does not imply that those nodes are not related. This poses an additional constraint on the learning objective of the node embedding in our case and standard node embedding methods (Grover and Leskovec 2016, Perozzi et al 2014) are not directly applicable.

BiNE (Gao et al 2018) is designed to utilize the heterogeneous nature of the network and can measure proximity even when the nodes are not connected by an edge, i.e. for two nodes of the same type. In our network, an *explicit relationship is depicted by an edge that connects a pair of patient and topic nodes*. On the other hand, *two patient nodes who are connected by an intermediate topic or two topic nodes connected similarly by a patient node are examples of implicit relationships*.

Similar to the BiNE model, we model the **relations** as the joint probability of two nodes. If  $v_i$  is the  $i^{th}$  patient node and  $t_j$  is the  $j^{th}$  then, we define the joint probability between two nodes by considering the local proximity between two nodes is given as:

$$P(v_i, t_j) = \frac{\alpha_{ij}}{\sum_{e_{ij} \in E} \alpha_{ij}}$$

where,  $\alpha_{ij}$  is the weight of the edge  $e_{ij}$ ,  $E$  denotes the edge.

The objective is to minimize the KL-divergence of the actual measure of the relationship ( $P$ ) and the expected value ( $P^\wedge$ ) computed from the vector representation  $t_j$  and  $v_i$ , represented as  $\omega_j$  and  $\omega_i$  respectively.

$$P^\wedge(v_i, t_j) = \frac{1}{1 + e^{-\omega_j^T \omega_i}}$$

Finally, the minimized KL-divergence to model the relation between the nodes of this network is given as:

$$\text{Minimize } O_1 = \text{KL}(P \parallel P^\wedge)$$

We now perform *random walks* on the obtained relation by using the Node2vec (Grover and Leskovec, 2016). This is represented in the network as a short walk from  $v_i$  to  $t_j$  via other patients and topics where,  $v$  represents patient node and  $t$  represents topic nodes. Node2Vec uses a mix of Breadth-First Search (BFS) and Depth-First Search (DFS). BFS uses importance of local neighbors or a micro-view, whereas DFS helps to obtain a more spread out connection with a macro-view (Grover and Leskovec, 2016). The node2vec model uses characteristics from both of these classic search models. Refer Figure 6.1.

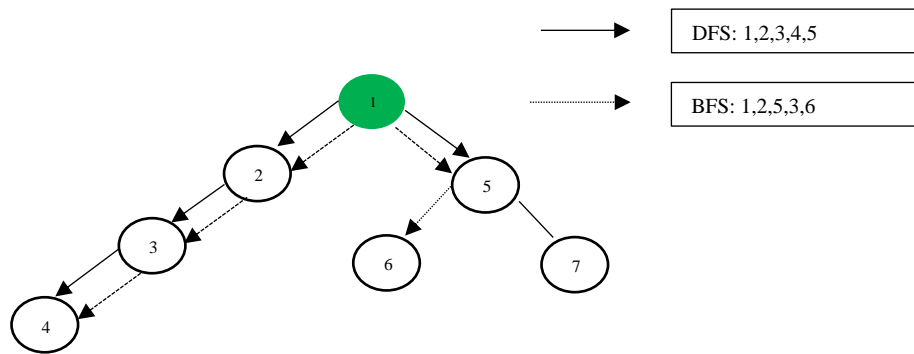


Figure 6.1: DFS and BFS for a *node* where number of walks = 4, where start node is 1

Node2vec model depends on 4 main hyperparameters:

- Number of walks: Number of random walks to be generated from each node in the graph
- Walk length: How many nodes are in each random walk
- P: Return hyperparameter
- Q: Inout hyperparameter
- Edge weights

P and Q are the probability that a node will retrace its path to the previous node or will go further to other undiscovered nodes, respectively. This probability depends on the edge weight ( $\alpha$ ), the normalized factor depends on the hyperparameters. Just like a word2vec skip-gram model where, for example, for the sentence "I love math", the probability of the word "math" depends on the occurrence of the words "I ", "like" i.e. its surrounding; a node2vec graph also generates these directed subgraphs for the nodes in a particular walk. Once we obtain the corpora of node sequences, we use the skipgram (Mikolov, 2013) technique to embed the nodes. With the design of the patient – topic network, we



aim to obtain embeddings that will give us the implicit as well as explicit relations as the nodes which are similar will be placed in closed proximity in the n-dimensional vector space. As per our patient – topic bipartite network, the sample of nodes  $A$  is a union of patient nodes  $V$  and topic nodes  $T$ . Hence,

$$A = V \cup T$$

## 6.4 Community Detection

The node embedding displays some level of organization at an intermediate scale (Zhan et al, 2017). At this mesoscopic level, it is possible to identify groups of nodes that are heavily connected among themselves, but sparsely connected to the rest of the network. These interconnected groups are called communities, or in other contexts modules, and occur in a wide variety of networked systems (Zhan et al, 2017).

Our ultimate goal is to identify unknown connections between patients and topics using the latent feature extraction. The purpose of identifying communities is to incorporate similar patients and topics into regions defined by a boundary. These bounded regions are used to discover the unknown relationships between two patients or two topics and even connecting patients with topics which were not directly implied by the data.

In our approach, we aim to create communities which include *both topics and patients* (users) and observe the proximity amongst topics and patients. Apart from observing users in the community who have no relation with other users or topics, it will be interesting to observe the diversity of topics in one community. If two topics do not fall under the same ‘Forum’ but are part of the same community, it can help us determine

the pattern and relationship between topics which would not have been easily detected just by observation.

For detecting the communities, we use *k-means* clustering and learn the clusters using Expectation Maximization (EM) algorithm. This method is divided into two steps - E and M steps. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster (Dabura, 2018). The objective of this method is to minimize cluster performance index, square-error and error criterion algorithm (Li and Wu, 2012). The algorithm tries to find K divisions to satisfy the optimal criterion. We manually went through a sample of clusters to observe whether our perceived outputs and the obtained outputs are similar in nature and modified the k-means model hyper-parameters accordingly (refer section 7.2.1).

## 6.5 Summary

In summary, we have extracted and obtained the patient and topic information from the data. In order to retain the original network structure but find the implicit as well as explicit relation between the patient and topic nodes, we will create our methodology as follows:

- Create a bipartite graph network by using patient and topic nodes and draw an edge between a patient and topic node if the patient has posted in that topic.
- To model the relationship from the bipartite network we apply the Bipartite Network Embedding or BiNE with random walk technique from the node2vec method to generate a corpora of node sequences.

- Next we use the skip-gram technique to create the node embeddings. The node embeddings are unique arrays of vectors that represent each node in an  $n$ -dimensional vector space where ' $n$ ' is predetermined.
- For the downstream evaluation tasks, we perform community detection to conduct the qualitative and quantitative analysis.

## CHAPTER 7

### PATIENT – TOPIC MODEL EVALUATION

#### 7.1 Overview

Through our model we wish to extract the latent features from a user – topic network to understand the social issues related to the disease state of a chronic disease. Due to the lack of standard datasets or related work with the exact same objectives, evaluating our method is challenging. We evaluate our method by building alternative models and demonstrate the improved performance of our method. We show that our design principles of building heterogeneous feature space (containing both patients and topics) has a greater value compared to a purely text-based method or a third alternative where only patient interaction is modeled.

We conducted several experiments to analyze the best configuration for the model, keeping in mind the large number of patient and topic nodes as well as computation time. After conducting a few experiments, we decided on a constraint to select the nodes from the original data to obtain the desired embeddings.

We conducted the evaluation for our embeddings in two phases:

1. Qualitative Analysis: Implement k-means clustering algorithm to obtain the communities and visualize the embeddings by using T-SNE to infer the results.
2. Quantitative Analysis: To understand the quality of the node-embeddings we created two baselines models: a. User Model

b. Text Based Model, to evaluate our method using topic coherence as well as used a reference data (same data) of topic clusters to evaluate the obtained embeddings and clusters.

In this section we describe the experiment setup for the modeling and discuss the inference and observations from the qualitative analysis as well as discuss the quantitative analysis conducted for understanding the quality of the obtained embeddings.

## 7.2 Experiment Setup

The dataset extracted from OBCF has over 94,000 users and around 4.4 million posts. From Figure 7.1 we get an overview of the extracted data better. We conducted several experiments to follow our thought process of linking the social issues associated with a disease which is breast cancer in our case, to the patients based on their current diagnosis. In one of the experiments we conducted, we constructed a heterogeneous network with patient nodes and the words they have used in their posts. Further we connected these two different types of node by drawing an edge between a user and the respective word if they have used that particular word in their posting. This method has its own set of challenges, as the same word can have different meanings in different contexts. For example, the word '*pain*' can be used in a surgical reference or used in a reference to the mental sufferings the patient is going through.

Finally, we created a model using the user – topic *bipartite graph network* (Figure 7.2) as we assume the topics in which the users post is the *summary of the context* being discussed in that thread.

This data consists of approximately 140,000 topics. For our user – topic model, to avoid over-fitting as well as underfitting of the model we chose users who have posted in more than 20 but less than 200 topics. We also eliminated the topics with less than 20 and more than 2000 users. This resulted in approximately 16431 individual users and 9242 topics for our bipartite graph.

Table 7.1: Count of users and topics after filtering

	<b>Users</b>	<b>Topics</b>
Original	94,000	140,000
New	<b>16431</b>	<b>9242</b>

We only used the User (patient) ID and no identity markers were used during network creation as well as modeling. In accordance with our assumption, we draw a connection between a user and a topic if and only if the user has posted in the respective topic (refer Figure 5.1) as, our aim is to find out whether there is a relationship between the users posting across multiple topics.

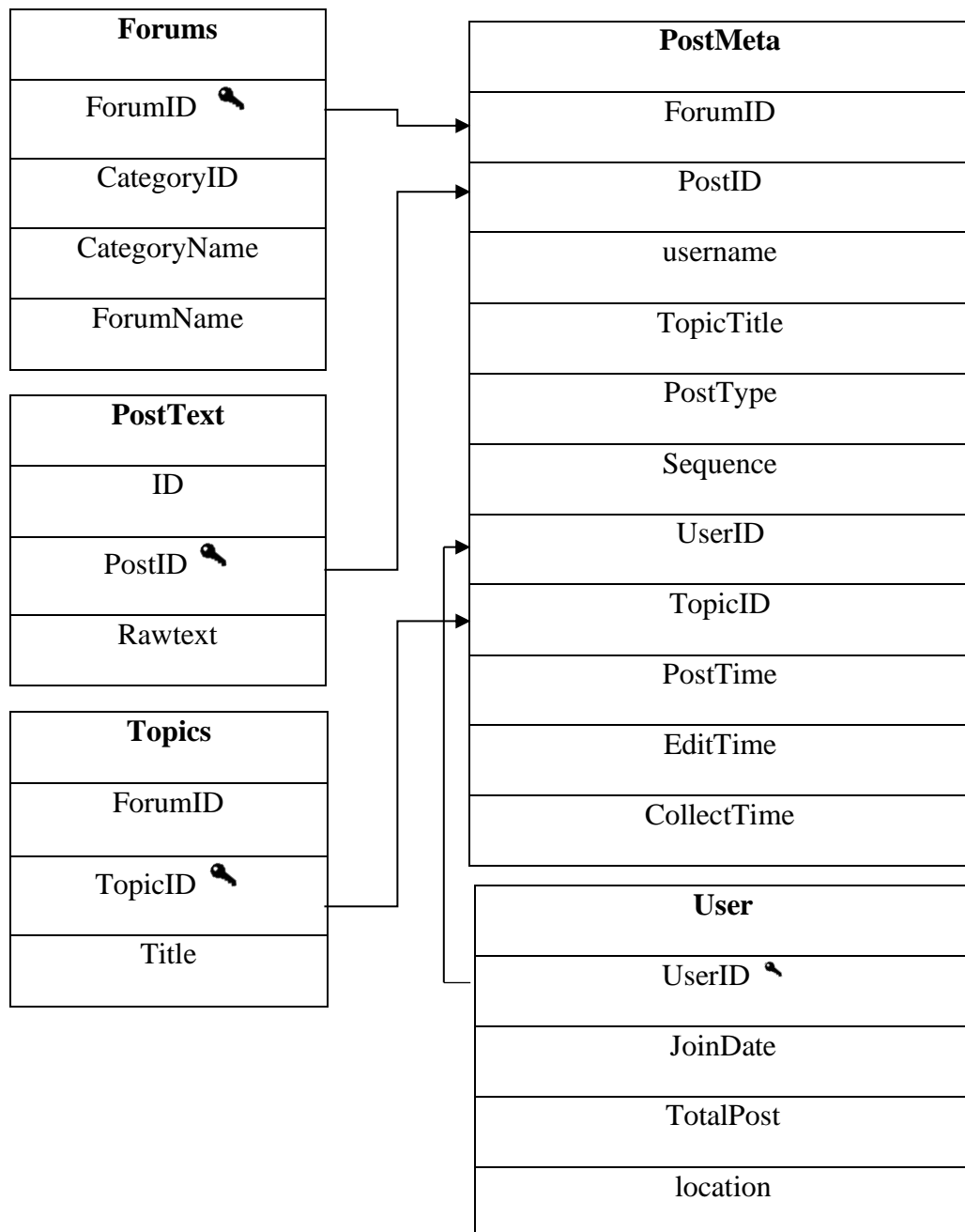


Figure 7.1: ERD for BRCA database

### 7.2.1 Hyperparameter Setup

As described in section 6.3, we perform the random walks using the node2vec model. The performance of a node2vec model depends on the values of its hyperparameters. Although the default value dimensions  $d$  is 128, Grover et al. (2016) observed that the performance tends to saturate once the dimensions of representations reach around 100. After conducting multiple simulation experiments and keeping in mind the large number of nodes in this model as well as the overall computation time, we decided empirically the value of number of random walks to be generated from each node ( $r$ ), the number of nodes in each random walk ( $l$ ) and the dimension ( $d$ ) of each node is as given in Table 7.2.  $p$  is the return hyperparameter and refer to the probability that the walk will retrace its path to a node that is previously visited. It is in the range (0,1) where a higher value indicates that it is very less likely for the walk to trace its previous path.  $q$  is inout hyperparameter which is the probability that the path will go either faraway into the network or remain closer to the original node. As per Perozzi et al (2014), a unity value of  $q$  will help the network maintain a balance or uniformity in the random walks to learn the d-dimensional feature representation.

Table 7.2: Hyperparameter for the embedding model

Parameter	Value
Dimension (d)	64
Walk Length (l)	30
No. of walks (r)	200
p	1.0
q	1.0



Once the hyperparameters are set, we will implement the node2vec algorithm on the patient – topic bipartite graph network to obtain the corpora of node sequences and then implement the skip-gram method to obtain the node embeddings. Thus, we will use these embeddings for downstream qualitative and quantitative analysis.

## **7.3 Qualitative Analysis**

### **7.3.1 Overview**

In this section we will observe and infer the clusters or communities obtained by implementing k-means algorithm on the patient – topic node embeddings. The primary aim of conducting qualitative analysis is to observe and understand the diversity of the nodes in an individual community. Diversity not only in terms of patients (user) nodes and topics but diversity in the topic forums i.e. topics discussing different context being associated together in the same community, users who have no common topic posts between them, being in close proximity i.e. in the same community. The thought behind conducting a qualitative analysis is to try and unearth the hidden connection between the nodes from the latent feature extraction method.

### **7.3.2 Node Embedding Clustering using k-means**

Andrey Bu, who has more than 5 years of machine learning experience and currently teaches people his skills, says that “*the objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a number ( $k$ ) of clusters in a dataset*” (Garbade, 2018 “*Understanding K-means Clustering in Machine Learning*”). Hence, we used k-means clustering

algorithm for community detection. As this is a first step towards a larger goal of creating a framework for the healthcare providers to create a holistic treatment plan for their patients, we use a hard boundary k-means clustering algorithm as we aim to understand the patient - topic and analyze the relation between nodes who are in close proximities.

We used the elbow method of K-means to find the optimum number of clusters with a sliding range of 500-1500 clusters. The rationale behind using this window was that each community should have around 20 nodes each for us to infer the relationships in that community. Our model has over 26,000 nodes hence the decision to use this sliding range. Through this method we obtained the total number of clusters to be created is 750. Figure 7.3 gives us a visual representation of the obtained communities from the node embeddings. We then manually browsed through two communities to understand the direct or indirect relations between the nodes in that community.

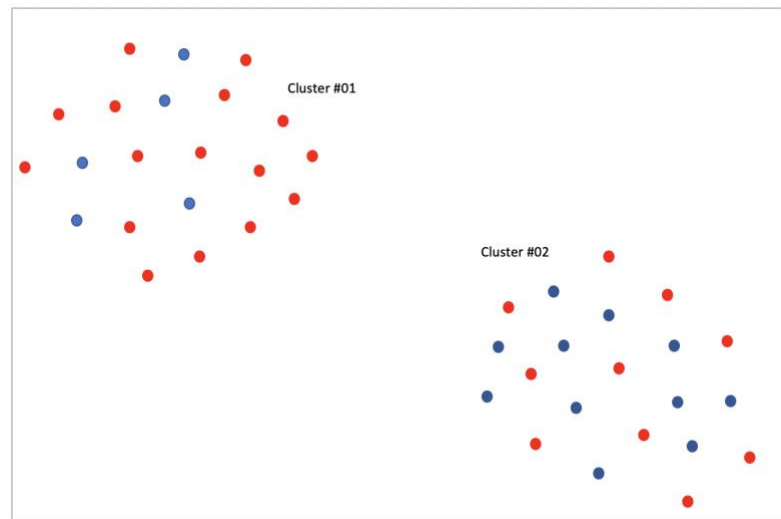


Figure 7.2: Representation of the obtained communities from the embedding data

We perform a hard – clustering algorithm - *k-means* as the first step towards obtaining and inferring the relationship between topics who are closely associated with each other, using patients (users) as intermediate variables. The obtained heterogeneous communities will give us an analysis of the topics that are *most closely* related to each other from the user’s perspective. Having the users and topics together in one community will help the healthcare providers directly link the users to the different issues or topics to which they are related to. Though we have created hard clusters for the purpose of this thesis, in the future we would like to extend this method further by creating a *dynamic machine learning solution*. This way we can not only associate the closest or most proximate nodes but also the nodes which have some relation to each other as the tiniest of associations can prove to be beneficial to someone’s disease management.

### **7.3.3 Community Detection: Observation and Inference**

In this sub-section we will manually go over two of the diverse communities obtained as a result of community detection algorithm performed over the obtained node embeddings. From the 750 communities, we observed that more 80% of the communities show diversity in terms of consisting of both patients as well as topics. This fulfills our intention of obtaining diverse communities. Each of these communities contain around 20 nodes obtained from the embeddings. This is the first time an analysis of this nature is being performed over this data using latent feature extraction embedding method hence, there is no existing model or labelled dataset to validate our obtained clusters or communities using traditional methods. Hence, the inference of these communities is performed per *our point of view* and needs to be clinically validated in the future.

### 7.3.3.1 Cluster #01

Table 7.3: Topics in Cluster #01

Topic ID	Topic Context
F6 T779992	Managing Side Effects Breast Cancer & treatment
F83 T773037	Not Diagnosed but Worried
F44 T758994	Breast Reconstruction
F69 T784857	Chemotherapy Before, During and After
F78 T775441	Hormonal Therapy Before, During and After

Table 7.4: Patients (users) in Cluster #01

Patients (Users)
V1
V2
.
.
.
.
V15

In “Cluster 01” consists of 5 topics (Table 7.3) and 15 users (Table 7.4). The 5 topics in this cluster fall under 5 different forums (Table 7.3). From this community we can observe the diversity in the topics right from “Not Diagnosed but worried” to “Chemotherapy Before, During and After”, “Breast Reconstruction” as well as “Managing Side Effects Breast Cancer & treatment” and “Hormonal Therapy- Before and After”. We can observe that there is some relation between these topics and as the

topic context (Table 7.3) suggests, they range from discussing chemotherapy and surgeries to the problems they faced due to the side-effects caused by them as well additional therapies required as a part of the treatment. This proximity between the topics can give the healthcare providers a broader view regarding their close association and helps them to understand the pattern and trend of the patient issues. Although these topics might seem unrelated from a naked-eye perspective, they are still highly correlated and need to be bounded together when it comes to the complete recovery of patients.

There are 15 users (patients) in this cluster (Table 7.4). Each user has posted in either of the topics contained in this group. It is interesting to note that no user has a common topic post between them from this cluster. From this we can infer that due to the nature of random walks the embeddings for the nodes in this cluster have a connection due to the latent features that are extracted from the bipartite network. Based on these associations between the patients (users), they can be advised on future difficulties and hurdles in-advance and the healthcare provider can also create the treatment plans accordingly.

### 7.3.3.2 Cluster #02

Table 7.5: Topics in Cluster #02

Topic ID	Topic Context
F91 T792393 + 2*	Surgery - Before and After
F96 T835504	IDC (Invasive Ductal Carcinoma)
F67 T796919 + 2*	Stage III Breast Cancer
F93 T784857	General Comments and Suggestions
F16 T776398	For Caregivers, Family, Friends & Supporters
F5 T793169	Just Diagnosed

Table 7.6: Patients (users) in Cluster #01

Patients (Users)
V1
V2
.
.
.
.
V9

In “Cluster 02’ there are 11 topics (Table 7.5) belonging to 6 different forums and 9 users (Table 7.6). In this cluster we observe that the topics in this community range from “Just Diagnosed” to “Stage III Breast Cancer” as well as “Surgery – Before and

After”. We can observe the diversity in the topic contexts in this community as well. The 10 topics in this forum belong to 6 different forums (refer Figure 4.1).

Topic F91 T792393, F91 T859005, F67 T796919, F93 T8605125 and F67

T27838556 has a common user (patient) say  $v^*$ , who has posted in it. User  $v^{**}$  has posted in topics F5 T788278 and F5 T793169. Both  $v^*$  and  $v^{**}$  are a part of this community.

As the patient  $v^*$  has had a diverse topic interaction, it will be interesting to observe the pattern of this patient’s posting and the issues this patient has discussed and map it to the other patients in this community as their close proximity in this feature space is indicating a relation between them.

Although through the qualitative analysis we can observe and infer the relationships between the nodes using a downstream analysis task which is community detection in our case, it is also important to analyze the quality of the obtained embeddings based on our technique of creating the patient - topic bipartite network.

## **7.4 Quantitative Analysis**

### **7.4.1 Overview**

In the previous section (Section 7.3), we observed and inferred the results of a downstream analysis task of community detection manually. Though we could make a sense of the obtained communities, we still need to understand the quality of this newly implemented path of creating a patient – topic bipartite network embeddings for latent feature extraction. For this we created two other baselines models: a. User Model  
b. Text Based Model, to evaluate our model. In this section, we will perform qualitative analysis of our node embeddings by comparing them with the embeddings of the other

two baselines models in two ways: a. Coherence b. Comparison with Labelled Data. In our case, we do not have a well-defined downstream task but as the final step, we perform community detection. We use the quality of the communities obtained to evaluate our overall methodology. Although our embedding framework is based on two variables – patients and topics, to be able to compare against the other baseline models, we only use the topic part of our model.

### **7.4.2 Baseline Models**

In this subsection we will see the construction and implementation of two baselines models we created to evaluate and validate our patient – topic bipartite network. Traditional models include either only textual data or only homogeneous node data as opposed to our method of constructing a heterogeneous network to address the issue of gaps in healthcare systems and unearth hidden links.

#### **7.4.2.1 User Model**

In order to unearth the latent connections between users of an online health forum, one of the ideas to implement the same is by creating a homogeneous user model. In this model we form a connection or draw an edge between two nodes between two users if the respective two users have posted in the same topic (Figure 7.4). This connection also gives an idea of the patients are linked to each other *unknowingly* through the online platform and how understanding the similarities between two patients can help either of them cope with their diagnosis better.



Once this user network is created, we implement the node2vec model to create the embeddings for the user nodes. Similar patients are placed in close proximity in the n-dimension vector (feature) space (Figure 7.5).

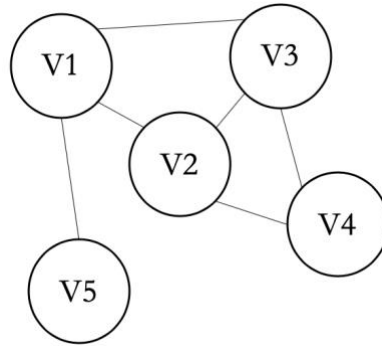


Figure 7.3: Network of users on online health platform

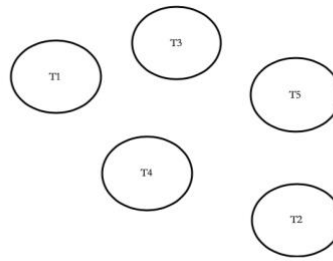


Figure 7.4: User model in representational vector space

The patient feature space is mapped into the topic space by replacing the patients with the topics they had directly interacted with. If more than one patient has participated in the topic, the final embedding for that topic will be the centroid of all the patients' vectors who have participated in that topic (Figure 7.6). For example, we have 3 users V1, V2 and V3 each with their embedding centroid to be C1, C2 and C3. If these 3 users have posted in Topic T1 then, the final embedding for T1 will be (C1, C2, C3). A minor drawback of creating this type of indirect topic network is that the final embedding for

the topic will depend on the order of the users who have posted in that respective topic (Figure 7.6).

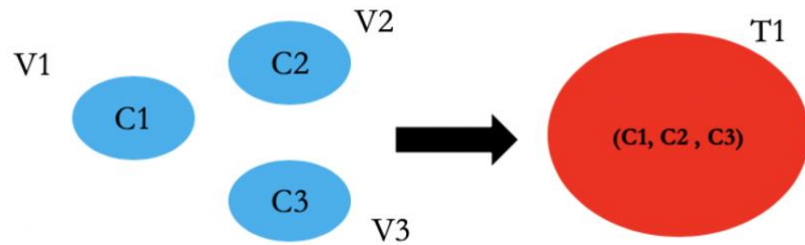


Figure 7.5: User embedding mapped to the topic space.

#### 7.4.2.2 Text Based Model

The second model is a text-based model. We extracted all the posts from the online health forum (OBCF) and implemented a word2vec model (Mikolov et al, 2013). A word2vec model Word2vec is a two-layer neural net that processes text by “vectorizing” words (Figure 7.7). Its input is a text corpus and its output are a set of vectors: feature vectors that represent words in that corpus (Li, 2019 “*A Beginner's Guide to Word Embedding with Gensim Word2Vec Model*”). A well-trained set of word vectors will place similar words close to each other in that space. The words paint, brush and palette cluster in one corner, while war, conflict and strife huddle together in another. This method is useful for understanding which words are in close proximity or in the same cluster in the feature space. This way we can observe which words are being used in a similar context and if these clusters give us an idea regarding the social issues faced by the breast cancer patients.

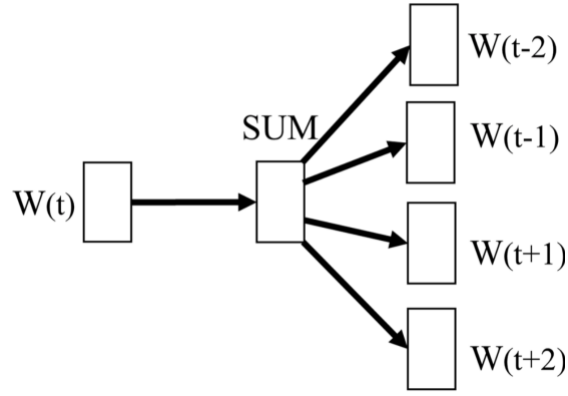


Figure 7.6: word2vec embedding using skip-gram (Jang et al, 2019).

### 7.4.3 Coherence

Topic Coherence is a measure that looks into the degree of similarity between items in the topic and it is often used to measure the quality of the vectors in embedding methods (Kharazmi and Kharazmi, 2017). These measurements help to understand how semantically interpretable the topics are. In our case, we extend this notion to measure the coherence of the embeddings of each node in the same community. There are numerous ways of measuring coherence statistically. Lau et al. (2014) showed that normalized pointwise mutual information (NPMI) showed the most consistent correlation with a manually annotated test set, compared to other metrics, such as other variations of pointwise mutual information, Log Conditional Probability (LCP) (Mimno et al, 2011) and pairwise distributional similarity (Aletras and Stevenson, 2013). We followed a similar approach and used NPMI to measure coherence in our case. By conducting a coherence evaluation, we aim to understand the degree of logical consistency of the

embeddings in a cluster. The higher the value the greater the relation between the nodes of that cluster.

Table 7.7: Comparing the performance of patient – topic network with baselines models using Normalized Pointwise Mutual Information (NPMI) to measure coherence

	<b>Coherence (NPMI)</b>
Patient – topic network model	<b>0.481</b>
User network model	0.237
Text – based model	0.294

From Table 7.7 we can observe that our patient – topic network model performs better than the other two baselines models when it comes to comparing the coherence of the embedded data. Our technique of creating and implementing a heterogeneous (bipartite) network consisting of both users and topics outperforms a homogeneous model consisting of only user data as well as a text-based model based on the posts made on OBCF.

#### **7.4.4 Comparison with reference dataset**

In the study conducted by authors Jones et al (2018), they categorized actionable topic clusters from individual forums of Breastcancer.org (section 4.3). Although this dataset was created to serve another purpose and does not exactly match our requirements, this was the only other reference dataset which had actionable topic

clusters from the respective OBCF. We used that manual set as a reference to evaluate our method using information retrieval evaluation metrics.

From section 7.2 and section 7.4.2 we have obtained embedding data for 3 different network models. As our reference dataset is for ‘topics’ we use only the *topic* nodes embedding data from our patient – topic network. Further, we perform a downstream community detection algorithm on three different embedding data to obtain clusters of close proximate topics. We conduct separate evaluation on each of the three distinct networks and their subsequently obtained communities. Each gives us different output and hence, we make use of the additional information about the topics (refer Figure 7.1) to infer the output obtained from the information retrieval metrics of precision and recall for understanding the topics situated in each cluster with respect to the topics in the clusters of the reference dataset.

Information retrieval systems are usually evaluated using two broad sets of metrics – online and offline (Manning et al, 2018). Online metrics measure the quality of the retrieved information using user engagement. In our case, we are evaluating our findings without the participation of any users, hence we need to use offline metrics. Among the numerous offline metrics, we selected precision-recall as we are not producing a ranked list of topics in this cluster. Thus, other metrics, such as, Discounted Cumulative Gain (DCG) or normalized DCG (NDCG) are not applicable here as these metrics focus returning a ranked lists and serves a better purpose for recommendation downstream tasks. Thus, we used precision-recall to measure the performance.

Table 7.8: Comparing the performance of patient – topic network with baselines models with respect to reference dataset in identifying communities.

	<b>Precision</b>	<b>Recall</b>
Patient – topic network model	<b>0.643</b>	<b>0.588</b>
User network model	0.428	0.314
Text – based model	0.507	0.422

From Table 7.8 we can note that, upon comparison with a reference dataset, our patient – topic model performs better than the other two baseline models in identifying communities with respect to the reference dataset. Higher precision – recall values indicate that our method has performed better than the other two to identify communities from the embedded data. This shows that here as well our technique showed higher performance as compared to other traditional methods.

From Table 7.7 and 7.8 we observe that the patient-topic network has performed better than the other variants or other two baseline models. This demonstrates the strength of the heterogeneous network we used and the ability of our technique to unearth more information. The two variables in the network were able to setup a channel for better interaction and be able to extract richer features from the underlying data.

## 7.5 Summary

In this chapter, we have discussed about the experiment setup as well as the qualitative and quantitative analysis which was performed to evaluate our patient – topic network. The entire setup of the experiment and the subsequent hyperparameters has been described. In the other sections of the chapter we have elaborated on the different types of evaluation performed to evaluate our patient-topic network embedding technique. We observed and inferred the communities we obtained from the downstream task of community detection that we performed on the embedded data and manually analyzed two of those clusters. Our quantitative analysis and evaluation proved the better performance of our model over the other two baselines we created in order to quantify our technique. The creation of the other two baseline models is described in detail in this chapter as well.

Thus, this thesis proves to be a successful step taken towards linking BC patients with the social issues they have to face as a consequence of their diagnosis and treatment. This work has a wide scope and can be further expanded to be integrated into the healthcare systems.

## CHAPTER 8

### CONCLUSION

This thesis is the first step towards bridging the gap between healthcare providers and consumers by creating a machine learning solution using the user-generated data extracted from online health platforms. The main objective of this thesis was to address the social issues faced by the patients suffering from a chronic disease by creating a network linking the users to the topics in which they have posted. This approach of keeping intact the original structure of data and inferring the clusters of proximate users and topics from this network will help the healthcare providers as well as the patients get a candid understanding of the issues that the patients are facing and this knowledge will hence be beneficial for inculcating appropriate steps in the disease management for other similar patients. Patients often refrain from sharing their problems with their healthcare providers or the healthcare providers overlook these issues or not communicating the diagnosis and the side-effects that come with it is a persisting issue. Our novel approach will help the healthcare providers get an understanding of these issues from the patients' perspective and help them to create holistic treatment plans.

As this issue is a long persisting one, different information systems have come up with various ways to shorten the gap between the healthcare providers and the consumers in involved. As discussed in *chapter 2*, there have been platforms developed to have a wholesome healthcare system and ensure that the patient's history is available at all times to respective healthcare personnel. Multiple case studies involving patients as well as providers have been carried out to understand the reality between the relationship between them.



While we address the existing issue in *chapter 1*, we study some of the existing platforms and case studies that have been conducted till now in *chapter 2*. In the subsequent chapters we discuss how machine learning and social network analysis help us to implement a model that will address this issue from a different perspective. In *chapter 3* we studied multiple works that have been previously conducted in the field of text mining, healthcare and breast cancer to design a novel approach for our cause.

Our approach was also based on the idea that we wish to bridge this gap based on data that was in its most candid form. No pressure to write the “correct” answer, no pressure of being judged or feeling low about oneself hence, we mined data from online health forum (OBCF) - Breastcancer.org, which serves as a platform for breast cancer patients and survivors to express themselves and discuss their problems as we studied in *chapter 4*.

In *chapter 5*, we analyzed this data to design our model to serve our purpose of connecting the social issues associated with a patient and his disease state based on topics he/she as well as other similar patients discuss on the platform. We created a bipartite (heterogeneous) graph network from patients and the topics in which they have posted to try and test our theory of understanding and unearthing the hidden links between users of an online platform as well as the context of their discussion (which is essentially the topics). For example, if two patients  $v_1$  and  $v_2$  have shared their experience on a large number of common topics  $T_i$  and  $v_1$  has also talked about other topics  $T_{i^*}$  (i.e.,  $T_{i^*} \cap T_i = \emptyset$ ), our model is able to connect  $v_2$  with  $T_{i^*}$ , as there is a connection between  $v_2$  and  $T_{i^*}$  via  $v_1$ .

To achieve this feat, we created a bipartite graph network using the BiNE concept of heterogeneous graph as we have two different nodes in our graph namely: a. patients (users) b. topics. As mentioned in *chapter 6*, we implement a node2vec algorithm to generate the random walks and the skipgram method to create the desired node embeddings from this network graph. These node embeddings are basically an array of vectors that represent each node on a n-dimension representational *feature space* on which we perform community detection as a downstream task for evaluation and validation purposes.

Although the best way to evaluate this work is through clinical validation, we performed a coherence model on the obtained communities to understand the degree of logical consistency in each cluster as well as used a reference dataset for information retrieval and classification on the obtained outputs. Taking into consideration the complexity of evaluating embedding data, we perform our analysis on the node embeddings in two ways as mentioned in *chapter 7* and is summarized as follows:

- Qualitative analysis: We manually observed and inferred two of the obtained communities to understand and form a connection between the nodes in those clusters. Apart from unearthing the connections between users in a community, we also got an understanding of which topics are closely related due to the nature of the embeddings which would've otherwise been tagged as unrelated.
- Quantitative Analysis: As there is no existing dataset or method to quantify the obtained results directly, we created two baselines to understand the quality of the embeddings using our technique of creating a patient-topic heterogeneous network. We observe the superiority of our network over the other two models

and observe the importance of including the patient nodes as variables in the network. We evaluate our model against the two other baseline models by showing the coherence of the new relationships and better performance compared to other similar methods.

In summary, we can claim that this approach has been effectively created as it retains the original structure of the network as well as exhibits a moderately good performance in the evaluations. However, there are certain *limitations* in the present approach. Even after proving the better performance of our model against two other models, the question of data security still persists and may raise privacy concerns. As this approach is constructed over the basis of user-generated data, we need to take into account the privacy rules set by the healthcare systems while extracting and creating our machine learning models. Our approach however minimizes the privacy risks as we represent the patient and their state as a vector which is then presented as their aggregated information. This vector, instead of representing the patient as an individual, will represent them with respect to other similar patients and the topics that are close to that patient. On the other hand, being able to place the patient within a larger perspective, can provide the healthcare providers with firsthand information about the issues from the patients (user) perspective to create holistic treatment and disease management plans. There is also no existing gold standard dataset or model that will help us evaluate our approach, model and obtained outputs in a straightforward manner. Thus, we needed to take indirect approaches to evaluate our outputs. Conducting a clinical study will help ascertain the obtained findings and perform appropriate analysis.

## 8.1 Contributions

The main contribution of this thesis is the creation of a novel approach to bridge the gap between healthcare providers and consumers based on the patient (user) postings on an online health platform. This thesis also presents as a first step of creating a web-based platform using text mining and dynamic machine learning. This information and understanding of foreseeable complications will enable the patients to equip better for the obstacles they have to deal with as well as the changes they need to imbibe as a part of their disease management. This will also help healthcare providers create a holistic treatment plan. The other major contributions include:

- Using text mining and machine learning to bridge the gap between healthcare providers and patients in an uncontrolled environment.
- This thesis also emphasizes on the importance of analyzing social media health platforms to complete the healthcare ecosystem by taking into consideration existing patients experiences. This is elaborated in section 4.1.
- Creation of a network graph by using data mined from an online social media platform connecting users and topics, where topics refers to the overview of the context being discussed in that thread. This is described in Section 4.1.
- Using the BiNE concept to create heterogeneous graphs to obtain communities consisting of both patients and topics, to understand the proximate distance between these nodes as mentioned in section 6.2.
- Through our embeddings and downstream community detection we can connect the topics who are different contextually but are actually related to each other

when it comes to the journey of the breast cancer patient and their complete recovery. This is discussed in section 7.3.

- Our technique of creating a patient – topic network embedding technique also preserves the structure of the network as well as identifying new relationships connecting similar patients or patients with similar topics, even when these relationships are not explicitly depicted in the data.
- Through our approach we can connect social issues that a patient will face based on their current diagnosis.

The following is also a relevant contribution of the present thesis:

- The chapter 2 of this thesis presents a detailed overview about the some of the platforms/tools/apps that have been already created to bridge the gap between healthcare providers and consumers, along with case studies that have been conducted by surveying patients suffering from chronic disease to understand their interactions with the healthcare system as well as their perception of it.

## **8.2 Future Work**

The current work has been evaluated in two ways, first a qualitative analysis whose results have been inferred from our point-of-view and second, quantitative analysis was conducted by creating two other baselines models where we proved that the patient-topic network technique performs better as compared to the other two models but this novel approach of creating a patient – topic network and the obtained communities needs to be validated by clinicians. A clinical validation will further strengthen our argument of creating the said patient – topic bipartite network to unearth the hidden

implicit as well as explicit relations on an online health platform.

This thesis is the first step towards creating a knowledge base of the social and personal problems related to BC using data and text mining with machine learning. Although this current technique will help preserve the network and help us gain insights about the network, we can build over this network by implementing additional text mining and natural language processing models that can help create a knowledge base containing both the medical as well as personal, professional and social issues associated with BC.

We have extracted the data from an online social media health platform where users “post” their experiences and opinions. Hence, the patient-topic network can be further strengthened by including the words from the posts that the users have used in the respective topic and use them as additional features. In natural language processing – machine learning models, we can establish a strong connection between two users if they have used common words. The usage of common keywords indicates a possibility of them undergoing a similar ordeal and having advanced knowledge of the experiences of one can help the other be better prepared to deal with the situations.

The acknowledgement and understanding of the social issues that the respective chronic disease patients face can be further integrated with the healthcare providers existing workflow. In order to create a complete healthcare system, the patients need to be aware about the issues that come with the diagnosis and the providers also need to take charge of making their patients aware about the same as well as make additional referrals in case it’s needed. As observed from the case studies in *section 2.3*, doctor-patient trust level needs to increase for the patients to follow and complete the treatment plan and that

will happen when the healthcare providers take into consideration their social issues and create a holistic plan. This can be facilitated or integrated into their workflow through our method.

One of the best ways to test our theory and technique is to create a web-based platform to deploy across various medical facilities in the state of Indiana and beyond to analyze the benefits of this study in practical.

Through our technique of creating embeddings and the nature of the network we can also implement a downstream task of creating automated timeline generation for each patient based on their current diagnosis. This roadmap will point out at the issues a patient faces at different points during and after the treatment.

We can analyze other social media health platforms for different chronic diseases, with a similar structure to OBCF and further extend this study to them as well and create a larger knowledge database to integrate with the healthcare system.

## PUBLICATION

### Publication from this thesis

- **Maitreyi Mokashi**, Enming Zhang, Josette Jones, Sunandan Chakraborty. 2020. Extracting Features from Online Forums to Meet Social Needs of Breast Cancer Patients. In ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '20), June 15–17, 2020, , Ecuador. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3378393.3403652>



## REFERENCES

- About Chronic Diseases. (2019, October 23). Retrieved June 22, 2020, from <https://www.cdc.gov/chronicdisease/about/index.htm>
- Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. 855–864.
- Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task. 43–47.
- American Cancer Society, 2020, “How Common Is Breast Cancer?: Breast Cancer Statistics.” [www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html](http://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html).
- Asch, D. A., Rader, D. J., & Merchant, R. M. (2015). Mining the social mediome. *Trends in molecular medicine*, 21(9), 528–529.  
<https://doi.org/10.1016/j.molmed.2015.06.004>

Bridging clinical information and communication gaps for better-informed care.

(2020, June 03). Retrieved July 06, 2020, from

<https://www.healthcareitnews.com/news/asia-pacific/bridging-clinical-information-and-communication-gaps-better-informed-care>

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 701–710.

Bull SA, Hu XH, Hunkeler EM, et al. Discontinuation of use and switching of antidepressants: influence of patient-physician communication. JAMA. 2002;288(11):1403-1409. doi:10.1001/jama.288.11.1403

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Cambridge university press.

Danielle H Bodicoat, Minouk J Schoemaker, Michael E Jones, Emily McFadden, James Griffin, Alan Ashworth, and Anthony J Swerdlow. 2020. Correction to: Timing of pubertal stages and breast cancer risk: The Breakthrough Generations Study. Breast Cancer Research 22, 1 (2020), 1–2.

- David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 262–272.
- Dina A Ragab, Maha Sharkas, Stephen Marshall, and Jinchang Ren. 2019. Breast cancer detection using deep convolutional neural networks and support vector machines. *PeerJ* 7 (2019), e6201.
- Dongdong Wang, Nayden G Naydenov, Mikhail G Dozmorov, Jennifer E Koblinski, and Andrei I Ivanov. 2020. Anillin regulates breast cancer cell migration, growth, and metastasis by non-canonical mechanisms involving control of cell stemness and differentiation. *Breast Cancer Research* 22, 1 (2020), 1–19.
- Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 1568–1576.
- Fang Hu, Jia Liu, Lihuan Li, and Jun Liang. 2019. Community detection in complex networks using Node2vec with spectral clustering. *Physica A: Statistical Mechanics and its Applications* (2019), 123633.

Francis, V., Korsch, B. M., & Morris, M. J. (1969). Gaps in doctor-patient communication: Patients' response to medical advice. *New England Journal of Medicine*, 280(10), 535-540.

Francois Modave, Yunpeng Zhao, Janice Krieger, Zhe He, Yi Guo, Jinhai Huo, Mattia Prosperi, and Jiang Bian. 2019. Understanding Perceptions and Attitudes in Breast Cancer Discussions on Twitter. *Studies in health technology and informatics* 2019 (08 2019).  
<https://doi.org/10.3233/SHTI190435>

Garbade, Dr. Michael J. "Understanding K-Means Clustering in Machine Learning." Medium, Towards Data Science, 12 Sept. 2018, [towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1](https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1).

George Palma, MD. "Electronic Health Records: The Good, the Bad and the Ugly. George Palma, MD, Medical Director, of Simpler Consulting, Discusses Benefits and Draw Backs for Electronic Medical Records." *Becker's Hospital Review*, 2020, [www.beckershospitalreview.com/healthcare-information-technology/electronic-health-records-the-good-the-bad-and-the-ugly.html](http://www.beckershospitalreview.com/healthcare-information-technology/electronic-health-records-the-good-the-bad-and-the-ugly.html).

Giulio Rossetti, Michele Berlingerio, and Fosca Giannotti. 2011. Scalable link prediction on multidimensional networks. In 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 979–986.

Goudge, J., Gilson, L., Russell, S., Gumede, T., & Mills, A. (2009). Affordability, availability and acceptability barriers to health care for the chronically ill: longitudinal case studies from South Africa. BMC health services research, 9(1), 75.

Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, and Mohammed Faisal Nagi. 2019. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. Journal of Healthcare Engineering 2019 (2019). <https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html> IEEE, 105-109

Immad Dabbura. 2018. K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. <https://towardsdatascience.com/k-meansclustering-algorithmapplications\protect\discretionary{\char\hyphenchar\font}{}{}evaluation-methods-and-drawbacks-aa03e644b48a>.

Ivedix, Author: IVEDIX Website: <http://test.ivedix.com> @iVEDiX, IVEDIX, A., & [Http://test.ivedix.com](http://test.ivedix.com), W. (n.d.). Mobile Healthcare Bridges Providers-Patient

Gap. Retrieved July 06, 2020, from <https://ivedix.com/mobile-healthcare-bridges-providers-patient-gap/>

Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. PloS one, 14(8), e0220976. <https://doi.org/10.1371/journal.pone.0220976>

Jennifer Y Sheng, Kala Visvanathan, Elissa Thorner, and Antonio C Wolff. 2019. Breast cancer survivorship care beyond local and systemic therapy. The Breast 48 (2019), S103–S109.

Jeremy A Greene, Niteesh K Choudhry, Elaine Kilabuk, and William H Shrank. 2011. Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. Journal of general internal medicine 26, 3 (2011), 287–292.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 530–539.

Josette Jones, Meeta Pradhan, Masoud Hosseini, Anand Kulanthaivel, and Mahmood Hosseini. 2018. Novel Approach to Cluster Patient-Generated Data Into

Actionable Topics: Case Study of a Web-Based Breast Cancer Forum. *JMIR medical informatics* 6, 4, e45.

Jungsik Park and Young Uk Ryu. 2014. Online discourse on fibromyalgia: text mining to identify clinical distinction and patient concerns. *Medical science monitor: international medical journal of experimental and clinical research* 20 (2014), 1858.

Keyuan Jiang and Yujing Zheng. 2013. Mining twitter data for potential drug effects. In *International conference on advanced data mining and applications*. Springer, 434–443.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal* 13 (2015), 8–17.

Laura Nyblade, Melissa A Stockton, Kayla Giger, Virginia Bond, Maria L Ekstrand, Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports* 9, 1 (2019), 1–12.

- Li, Zhi. “A Beginner's Guide to Word Embedding with Gensim Word2Vec Model.” *Medium*, Towards Data Science, 1 June 2019, [towardsdatascience.com/a-beginners-guide-to-word-embedding-with-gensim-word2vec-model-5970fa56cc92](https://towardsdatascience.com/a-beginners-guide-to-word-embedding-with-gensim-word2vec-model-5970fa56cc92).
- Lori J Goldstein, Raymond P Perez, Denise Yardley, Linda K Han, James M Reuben, Hui Gao, Susan McCanna, Beth Butler, Pier Adelchi Ruffini, Yi Liu, et al. 2020. A window-of-opportunity trial of the CXCR1/2 inhibitor reparixin in operable HER-2-negative breast cancer. *Breast Cancer Research* 22, 1 (2020), 1–9.
- McCallum, Andrew Kachites. *MALLET Homepage*, 2002, [mallet.cs.umass.edu/](http://mallet.cs.umass.edu/).
- Ming Gao, Leihui Chen, Xiangnan He, and Aoying Zhou. 2018. Bine: Bipartite network embedding. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 715–724.
- Mira, J. J., Guilabert, M., Pérez-Jover, V., & Lorenzo, S. (2014). Barriers for an effective communication around clinical decision making: an analysis of the gaps between doctors' and patients' point of view. *Health Expectations*, 17(6), 826-839.



- Mohamad Abdolahi Kharazmi and Morteza Zahedi Kharazmi. 2017. Text coherence new method using word2vec sentence vectors and most likely n-grams. In 2017 3<sup>rd</sup> Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS).
- Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers. 13–22.
- Neuner-Jehle, S., Zechmann, S., Maissen, D. G., Rosemann, T., & Senn, O. (2017). Patient–provider concordance in the perception of illness and disease: a cross-sectional study among multimorbid patients and their general practitioners in switzerland. *Patient preference and adherence*, 11, 1451.
- Robert A Hiatt and Nancy Breen. 2008. The social determinants of cancer: a challenge for transdisciplinary science. *American journal of preventive medicine* 35, 2 (2008), S141–S150.
- Roger Mc Lean, Ellen MH Mitchell, E Nelson La Ron, Jaime C Sapag, Taweessap Shao-dian Zhang, Edouard Grave, Elizabeth Sklar, and Noémie Elhadad. 2017. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *Journal of biomedical informatics* 69 (2017), 1–9.

Siraprapasiri, et al. 2019. Stigma in health facilities: why it matters and how we can change it. BMC medicine 17, 1, 25.

Stacy M Carter, L Claire Hooker, and Heather M Davey. 2009. Writing social determinants into and out of cancer control: an assessment of policy practice. Social science & medicine 68, 8 (2009), 1448–1455.

Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ. 1995;152(9):1423-1433.

Thakkar, S. (2018, October 09). Bridging the Communication Gaps Between Patients and Providers. Retrieved July 06, 2020, from <https://myvetahealth.com/bridging-communication-gaps-patients-providers/>

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).

William B Grant. 2020. Lower vitamin D status may help explain why black women have a higher risk of invasive breast cancer than white women. Breast Cancer Research 1 (2020), 1–2.

Yang Yang, Nitesh Chawla, Yizhou Sun, and Jiawei Han. 2012. Predicting links in multi-relational and heterogeneous networks. In 2012 IEEE 12th international conference on data mining. IEEE, 755–764.

Yongcheng Zhan, Ruoran Liu, Qiudan Li, Scott James Leischow, and Daniel Dajun Zeng. 2017. Identifying topics for e-cigarette user-generated contents: a case study from multiple social media platforms. *Journal of medical Internet research* 19, 1 (2017), e24.

Yonguo Li and Haiyan Wu. 2012. A clustering method based on K-means algorithm. *Physics Procedia* 25 (2012), 1104–1109.

Z. Jin, R. Liu, Q. Li, D. D. Zeng, Y. Zhan, and L. Wang. 2016. Predicting user's multi interests with network embedding in health-related topics. In 2016 International Joint Conference on Neural Networks (IJCNN). 2568–2575.  
<https://doi.org/10.1109/IJCNN.2016.7727520>

# MAITREYI MOKASHI

## PROFILE

Master's in Applied Data Science  
Graduate student, seeking full-time job opportunities.

## CONTACT

### Phone

+1(317)-726-9466

### Email

maitreyimokashi@gmail.com

### LinkedIn

linkedin.com/in/maitreyimokashi

## PUBLICATIONS

**ACM COMPASS 2020:** A Novel Approach for Extracting Features from Online Forum to Meet Social Needs of Breast Cancer Patients

## CERTIFICATIONS

- HIPAA Training
- CITI Training
- DataCamp Python
- DataCamp Machine Learning

## INTERESTS

- Research
- Data Extraction
- Data Mining
- Data/Statistical Analysis
- Machine Learning
- Natural Language Processing
- Deep Learning
- Statistics
- Data Visualization
- Image Processing
- Database Management
- Artificial Intelligence
- Big Data
- Cloud Computing

## SKILLS

## EDUCATION

### Indiana University | Indianapolis, IN

AUGUST 2018 – AUGUST 2020

Master's in Applied Data Science

Core courses (till date) include: Database Design and Management, Machine Learning with Python, Deep Learning, Cloud Computing, Data Visualization, Web-Database Development, Statistical Inference and Data Analytics with R

### Symbiosis International University | India

AUGUST 2014 – MAY 2018

Bachelor of Technology in Electronics and Telecommunication Engineering

## WORK

### Indiana University | Indianapolis, IN

Graduate Research Assistant

AUGUST 2018 – PRESENT

**COMET LAB:** Working with Prof. Robert 'Skip' Comer to design and develop volunteer and patient management systems for IU-Medical School and Jane Pauley Community Health Center respectively. My work includes coordinating with clients, designing data-driven solutions, and developing web-database systems. I helped to conduct statistical analysis on the IU-School of Medicine Outreach Clinic data to observe the relation between intense academic semesters and the volunteering frequency of the students at the clinic. I have hands-on experience with HIPAA compliant biomedical/clinical databases as well as using REDCap application as result of collaborations between COMET Lab and various clinical research teams on and off the IUPUI campus.

Graduate Teaching Assistant

AUGUST 2019– DECEMBER 2019

Teaching Assistant for Applied Statistics for Biomedical Informatics (using R) course. Assisted Dr. Pradhan with classroom lecture materials, exams and project material. Helped to create and manage the infrastructure of the class, research, grade assignments and guiding students in their coursework.

### Cummins Inc. | Pune, India

JUNE 2017 – DECEMBER 2017

Operating Systems – Intern

Generated specialized automated doser patterns using Embedded C to control the automated injection of fluids in the exhaust system, reducing the greenhouse gas emissions in exhaust systems for engines and generators.

## MASTER'S THESIS

### A Novel Approach for Extracting Features from Online Forum to Meet Social Needs of Breast Cancer Patients

This study aims to create a framework that can connect personal and social issues associated with a disease, with a special focus on breast cancer. Treatments, drugs can have adverse effects on a patient's daily life. The ultimate goal is to connect a user's medical information with such non-medical information, that is mined from an online health forum, using Social Network Analysis, Data and Statistical Analysis and Natural Language Processing (NLP) along with Machine Learning models. This information when integrated with EHR, will help physicians review and acknowledge the possible external effects and factors in their treatment plan. *Paper Accepted: ACM COMPASS 2020.*

## SELECTED ACADEMIC PROJECTS

# MAITREYI MOKASHI

- SQL
- Python
- Machine Learning + libraries
- Natural Language Processing
- R
- NoSQL
- Deep Learning
- Spark/Hadoop
- LaTeX
- PHP
- HTML + CSS
- Java
- D3
- C++
- Shell Scripting (Linux/Bash)

## TOOLS/SOFTWARE

- MySQL
- Natural Language ToolKit (NLTK)
- Jupyter Notebook
- Python IDE
- R Studio
- Access
- TrackVis
- AWS
- Azure
- Databricks
- Visio
- Tableau
- Microsoft Access
- Codelgniter

## SOFT SKILLS

- LaTeX Documentation
- Project Coordination
- Presentations
- Student Mentoring
- Microsoft Word
- Microsoft PowerPoint
- Microsoft Excel

### **Forecasting Stocking of Narcan for Opioid Overdose Reversal by EMS in Indiana** | Data Forecasting

A time-series model to forecast trends and patterns in the EMS calls related to Narcan Administration in the different counties of the state of Indiana. We aim to find the most efficient model between LSTM, ARIMA and Prophet model.

### **Recommendation System using ALS model** | Cloud Computing

A book recommendation system based on the Alternating Least Square Algorithm using Hadoop for Cloud Computing. Implemented k-fold cross-validation for model evaluation.

### **Airport Management System** | Database Design and Development

Developed an airport management database system. Designed a data-book which includes Crow's foot relation, ERDs, ERM and a SQL query efficient system.

### **Admissions in the United States** | Tableau Platform

Tableau platform to visualize the admission to colleges in USA. The visualizations help to analyze various trends in admissions related to most preferred university, diversity, universities preferred by women, application to enrollment visualizations etc.

### **How do they feel about it? Drug Related Sentiment Analysis Pipeline** | Natural Language Processing for Biomedical Records and Scripts

Sentiment analysis of the various drugs (from BreastCancer.org) used in the Immunotherapy treatment (BC), extracted via UMLS meta-thesaurus mapping. Annotated the raw data set as positive, negative and neutral for implementing and evaluating the model. Compared these results with VADER sentiment analyzer to compare the performance of a pre-trained model with a model trained from manually annotated data.

### **Interactive Data Visualization Map** | JavaScript and D3.js

Interactive webpage for data visualization of Dr. Jon Snow's London's cholera outbreak data using D3 and JavaScript.

<https://maitreyi96.github.io/visualization/visualization/index.html>

### **Breast Cancer Prediction: IDC Histology Images** | Deep Learning

The aim of the project is to identify and classify the Histology images as positive or negative, according to the presence of Invasive Ductal Carcinoma (IDC) in the histopathology images of the breast tissue.

### **Automated Track Extraction to Observe the Deficit in an Illiterate Brain** | Tractography and ML – Undergraduate Final Year Project

Performing Tractography to identify a model brain to train and automate the track (brain tissue) extraction process using TrackVis from multiple scans and compare the density of brain fiber. This study aims to find a difference in the brain fiber structure between a literate and illiterate brain and to analyze whether receiving an education can help postpone Alzheimer's.



# Bridging the Gap between Healthcare Providers and Consumers: Extracting Features from Online Forum to Meet Social Needs of Patients Using Network Analysis and Embeddings

**Maitreyi Mokashi**

MS Applied Data Science

School of Informatics and Computing

Indiana University Purdue University Indianapolis

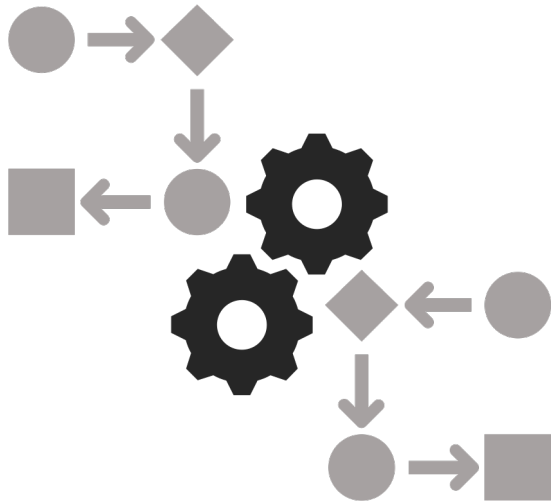
Committee: Sunandan Chakraborty, Ph.D. (Chair)

Josette Jones, Ph.D.

Jiaping Zheng, Ph.D.

Date: 07/27/2020

# Overview



1. Introduction
2. Breast Cancer: USA Stats and Facts
3. Bridging The Gap: Web-based tools
4. Bridging The Gap: Case Studies
5. Data
6. Goal
7. Problem Definition
8. Methodology
8. Evaluation and Results
9. Tools and Algorithm
10. Contributions
11. Future Work
12. Conclusion

# Introduction

“ Sometimes it’s harder to see the bigger picture and recognize that although the patient is coming to you for medical care, holistically they are dealing with a lot of other things.”

- *Fatima Haq, Global Medical Education Advisor at Eli Lilly and Company*



# Introduction

## Chronic Disease

Cancer

Heart Disease

Diabetes

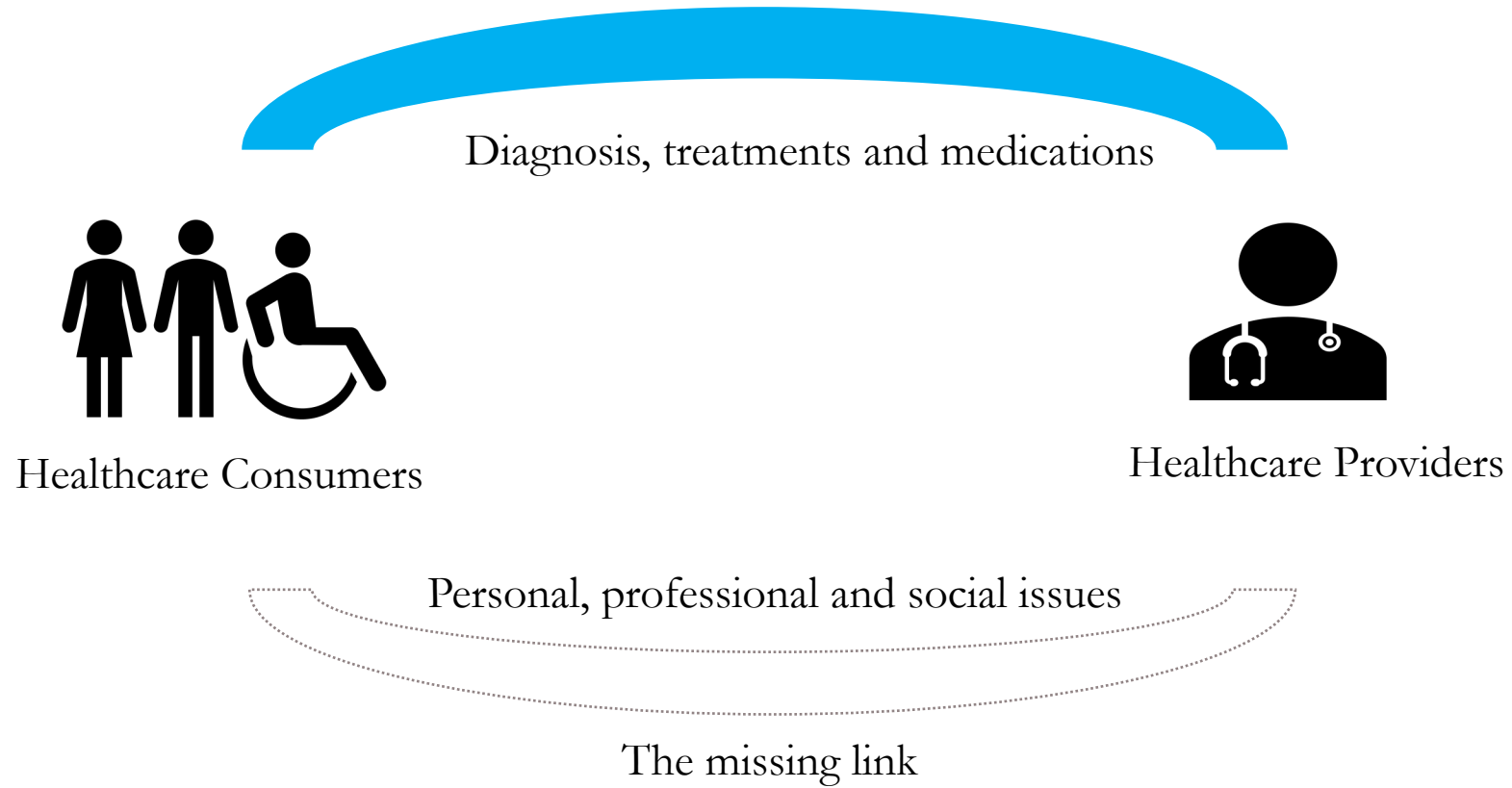
Death

Disability

**6 out of 10 American live with  
at least one chronic disease.**

(National Center for Chronic Disease  
Prevention and Health Promotion,  
2019)

# Introduction



# Breast Cancer : USA Stats and Facts



Breast Cancer is the most common type of cancer amongst American women next to skin cancer.



It is the second leading cause of death in women caused by cancer, the first being lung cancer.



The American Cancer Society approximates 276,480 newly diagnosed cases in 2020, affecting 1 in 8 women and 1 in 883 men with a prediction of almost 42,000 deaths.



The overall death rate due to cancer is decreasing by 1.3% per year due to early screening and treatment advances

The American Cancer Society's reported that as of January 2020, there are more than 3.5 million women with a history of breast cancer in the U.S. This includes women currently being treated and women who have finished treatment.

# Breast Cancer : Medical, Social and Personal Concerns

⊕ Some medical/treatment related concerns

What are the side-effects of the medications ?

What is the success rate of this clinical trial ?

Surgery?  
Should I go for it??

How much does it cost?



Some social/personal related concerns

Am I beautiful anymore ?  
Am I enough?

He wants a divorce?

I'm so young! Which guy will accept me, knowing my condition?

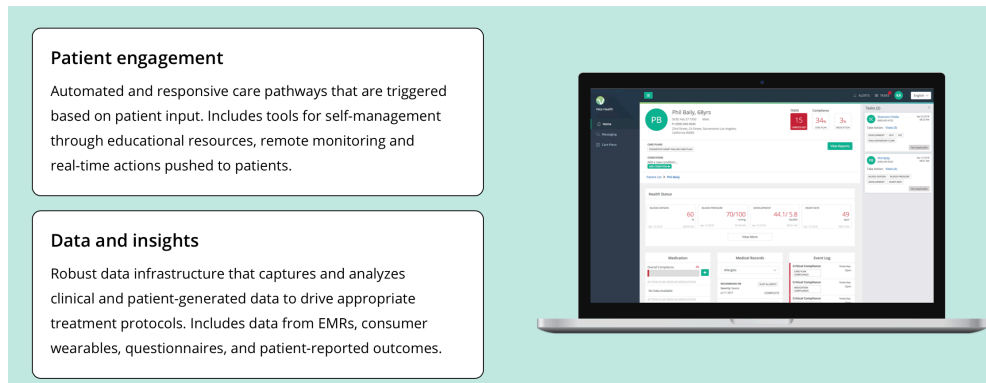
Medications do not leave sufficient funds for food.  
What will I feed my kids?



Snippets from Online Breast Cancer Forum

# Bridging The Gap : Web-based tools

1. The Ascom Myco 3 by ascom for healthcare [7]
2. Veta Health [8]
3. iVEDiX Health [11]



## Veta Health

1. Study 1: Effective patient provider communication can improve a patient's health as much as many drugs can (Stewart, 1995)<sup>[9]</sup>
2. Study 2: The study resulted in a positive correlation between effective communication and improved patient health outcomes (Bull et al, 2002)<sup>[10]</sup>

\* Screenshots from the website: <https://myvetahealth.com/platform/>

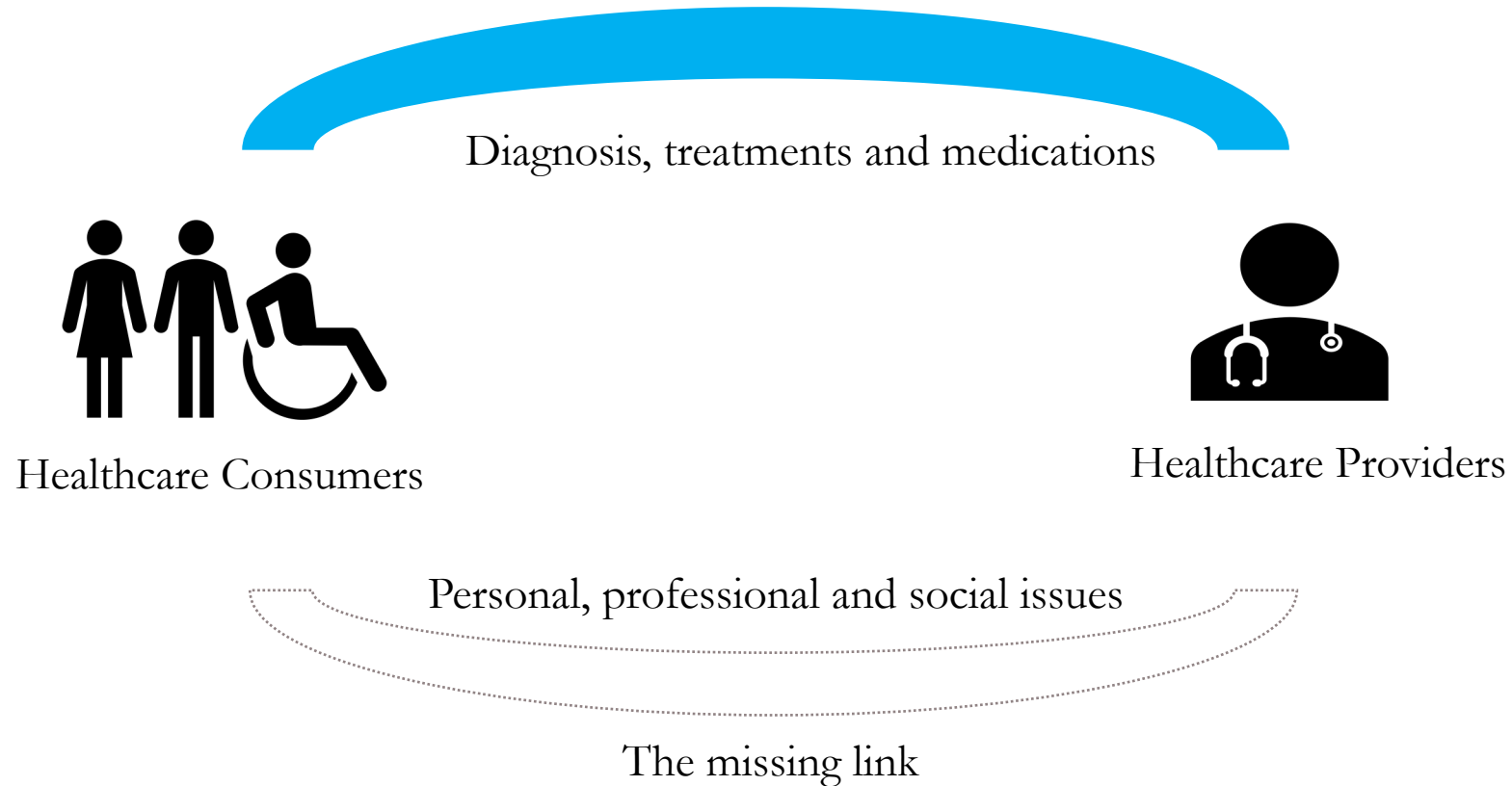
# Bridging The Gap : Case Studies

- Francis et al (1969) <sup>[12]</sup> “Gaps in doctor-patient communication: Patients' response to medical advice”
  - Objective: study of 800 out-patients visits to Children's Hospital of Los Angeles to explore the interaction between doctor and patient on patient satisfaction
  - Results: a lack of warmth in the interaction as well as failure to receive a proper explanation of their diagnosis and treatment became a major factor in the noncompliance from the patients end.
- Goudge et al (2009)<sup>[13]</sup> “Affordability, availability and acceptability barriers to health care for the chronically ill: longitudinal case studies from South Africa”
  - Objective: conducted a household survey of approximately 30 households over 10 months with descriptive narratives, that helped gain textual data to understand the interactions with the health system.
  - Results: Of the cases, 34 cases were chronic illness, only 21 (62%) cases had an allopathic diagnosis and only 12 (35%) were receiving regular treatment. Livelihoods exhausted from previous illness and death, low income, and limited social networks, prevented consultation with monthly expenditure for repeated consultations as high as 60% of income.

# Bridging The Gap : Case Studies

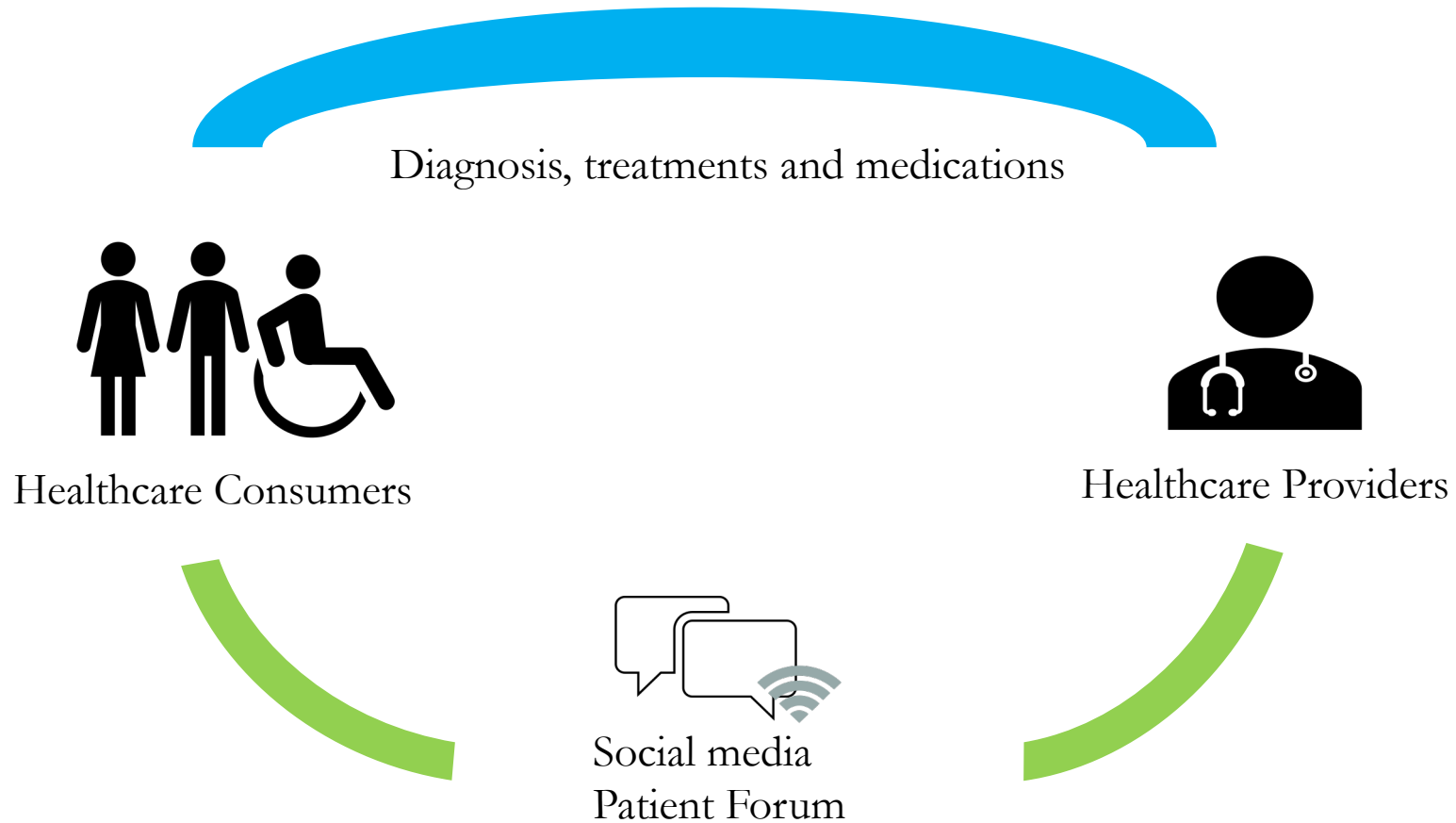
- Mira et al (2012)<sup>[14]</sup> “Barriers for an effective communication around clinical decision making: an analysis of the gaps between doctors' and patients' point of view.”
  - Objective: conducted a study in the 14 health centers belonging to 3 primary care districts and 3 hospitals in Spain. Their study included 764 patients and 327 physicians to determine whether patients consider the information obtained from the physicians enough.
  - Results: Their study showed that patients are not normally informed about medication interactions, precautions and foreseeable complications.
- Gu et al (2019)<sup>[15]</sup> “Development of a consumer health vocabulary by mining health forum texts based on word embedding: semiautomatic approach”
  - Objective: The vocabulary gap between consumers and professionals in the medical domain hinders information seeking and communication. The objective of this paper is to develop a method for identifying and adding new terms to consumer health vocabularies, so that it can keep up with the constantly evolving medical knowledge and language use.
  - Results: Their algorithm can correctly identify over 80% of the synonyms by just searching from the top 10 candidates of a certain medical term.

# Data: Thought Process





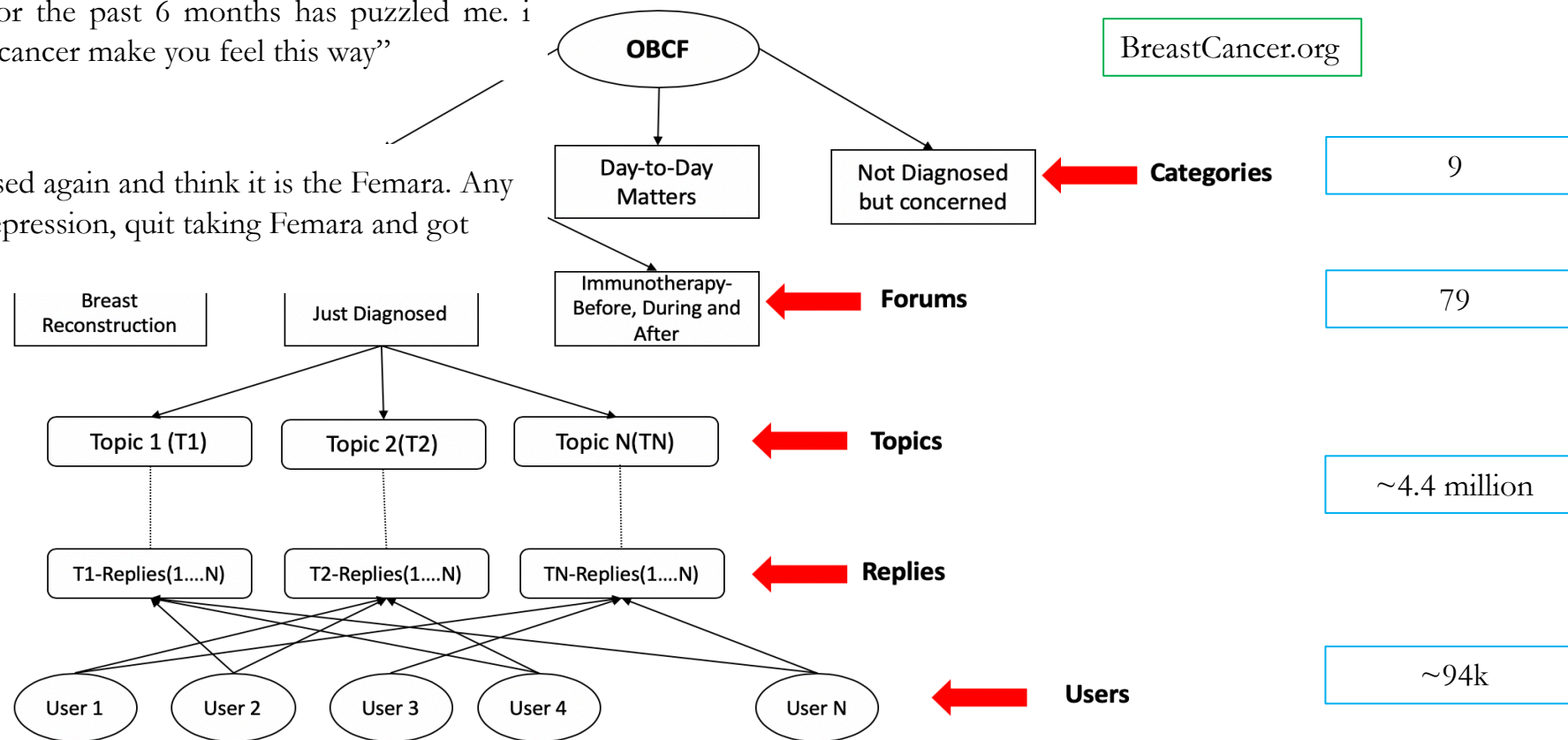
# Data: Thought Process



“before i even thought about the possibilty of having breast cancer, i had become very very tired and fatigued. have felt this way for the last 6 months. i have recently been diagnosed with breast cancer and will have surgery soon and chemo. but the fatigue and tiredness for the past 6 months has puzzled me. i sleep well. does having cancer make you feel this way”

I am getting so depressed again and think it is the Femara. Any of you out there had depression, quit taking Femara and got better?

# Data



# Data : Stats

<b>Posts per type</b>	<b>Forums</b>	<b>Topics</b>	<b>Users</b>
Max	616598	56091	48986
Min	11	1	1
Mean	56304	31	47

Table 1: Analysis of post in each level of OBCF

# Goal

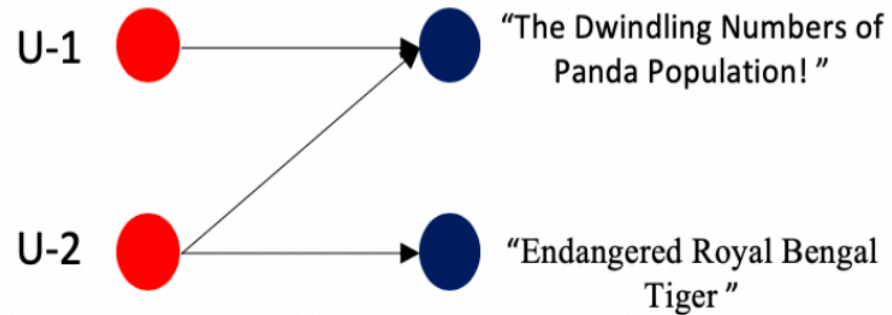


Our goal states that advance knowledge of such issues will help the physicians provide additional recommendations or referrals for the patients and provide them with a more holistic treatment.



We present a novel approach that will mine social/personal issues of a breast cancer patient along with their medical conditions from online health forum and represent them as latent features.

# Problem Definition



Network of Users - Topics

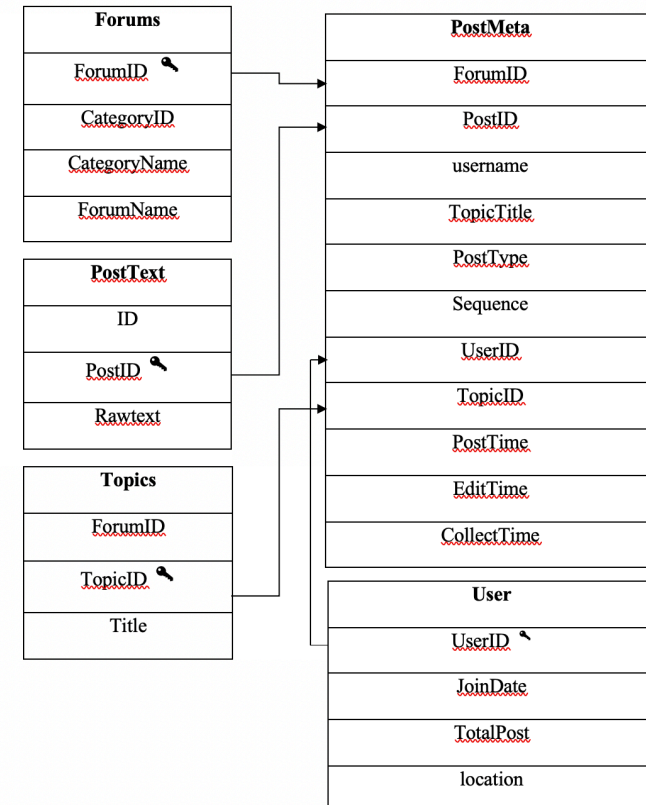


Latent Feature Space Representation

# Problem Definition

We formulate the problem with two known variables:

1. patients (users) ( $V$ )
2. topics ( $T$ )



ERD for the BC database

# Problem Definition

1. A network 'G' where,

$$G = (V, T, E)$$

2. A patient  $v_i$  and a topic  $t_j$  will be connected by an edge where,

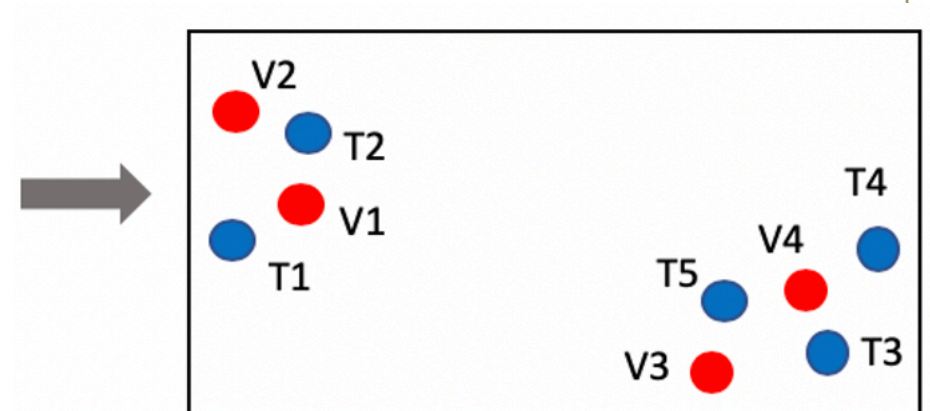
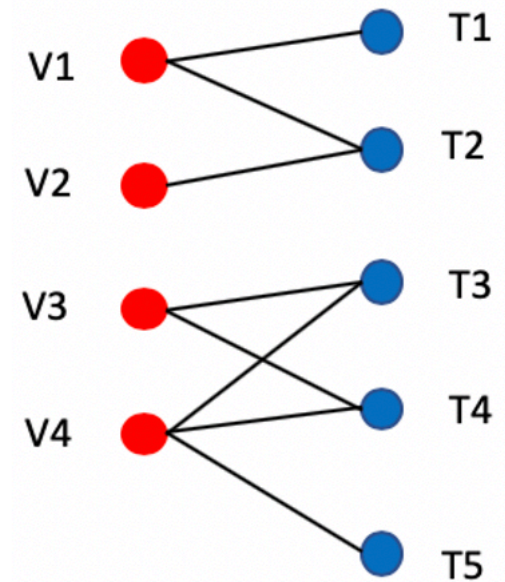
$$e_{ij} \in |V| \times |T|,$$

if the patient  $v_i$  have participated (posted/replied) in the topic  $t_j$

3. Our final goal is to design a mapping function  $\psi$ ,

$$\psi : V \cup T \rightarrow R^K,$$

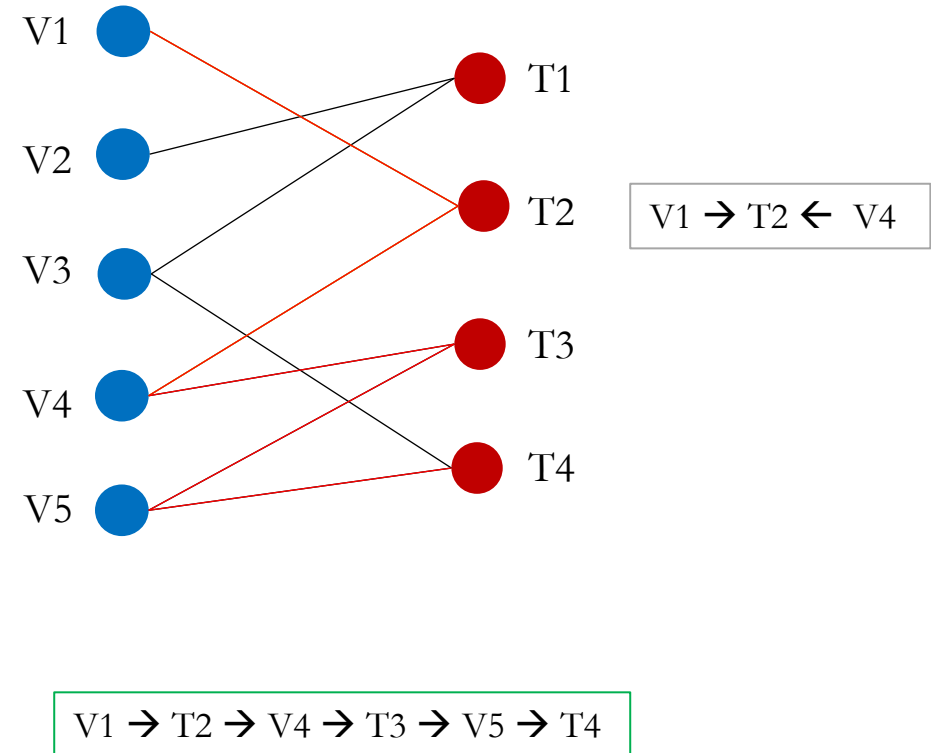
$K$  represents the dimension pre-determined of the feature space.



# Methodology

## Patient – Topic Network Embedding

- To preserve the patient - topic network we create a Bipartite Network
- To model the implicit and explicit relations, we use BiNE<sup>[1]</sup> theory of network embedding for bipartite network.
- We use Node2Vec<sup>[2]</sup> to generate the random walks and create the node embeddings
- We use Python, Machine Learning Libraries and Natural Language Processing for our modelling.

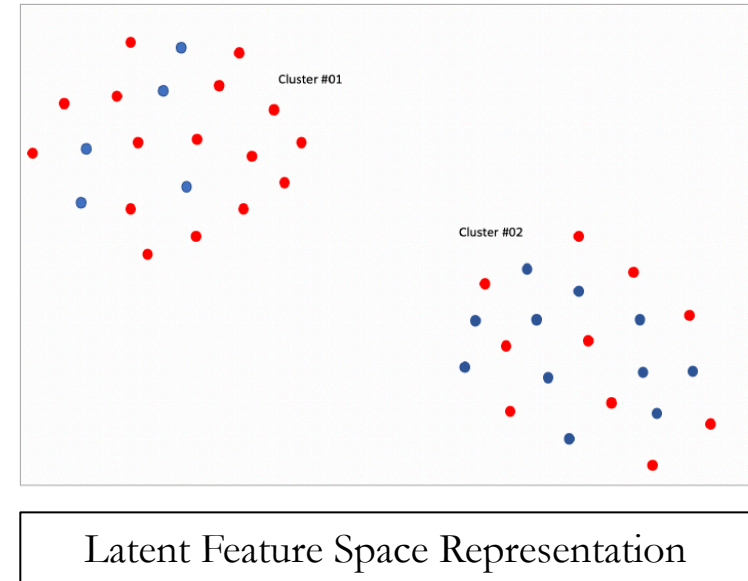




# Methodology

## Community Detection

- a. From the node embeddings, we can identify groups of heavily connected nodes which are sparsely connected to others in the network using k-means to find the number of clusters.
- b. These interconnected groups are called communities<sup>[3]</sup>.



# Evaluation and Results



Experiment Setup



Qualitative Analysis



Quantitative Analysis

# Experiment Setup

## Original Data

- a. No. of users: **94,393**
- b. No. of topics: **140k**

To avoid the possibility of over-fitting the model :

- We chose users who posted in more than 10 topics but less than 200 topics.
- If  $n$  is the number of topics a user has posted in then, we selected users :  **$10 < n < 200$**

## Experiment Data

- a. No. of users: **16431**
- b. No. of topics: **9242**

# Node2Vec: Hyperparameter Setup

Parameter	Value
Dimension (d)	64
Walk Length (l)	30
No. of walks (r)	200
p	1.0
q	1.0

Table 2: Hyperparameters for node2vec algorithm

# Node2Vec: Embeddings

```
11815 -1.1212362 -11.610572 0.27183175 10.8449335 -5.3459044 -2.6064928 -0.26061067 -4.5086575 -3.0102742 -8.91818 0.75674766 8.943197 1.1289129 -2.4953578 14.13245
0.94417053 5.42178 -5.2022223 -6.080257 -4.083056 -10.723732 9.454253 3.7048564 3.9062104 -6.691343 0.03680433 5.389365 -5.8425965 -12.698317 -5.501305 -6.274668
0.96091866 -0.5647932 6.926268 0.76904863 11.648858 -8.9495 5.8648505 4.1465907 4.7448945 5.322708 6.028748 -4.6033945 -4.712585 2.2667441 4.1041555 -20.851229
-3.052759 8.077602 6.1703854 2.9329443 -2.1257193 -3.2907267 2.585199 6.675979 -0.70538396 6.078326 -4.6551313 -4.2028546 4.621041 0.44258738 2.2070305 -1.7769442
2.6664495
F69T707348 -3.9646373 6.907322 4.277927 4.0467806 1.3467109 -2.5672004 -1.9465439 -1.767857 9.628621 0.7257606 1.6849649 -5.7741976 4.7679777 -1.1445857 4.1638703
9.596284 -6.4262977 1.8114915 -2.342598 -9.135571 -3.105016 -8.411502 -4.878233 -7.898892 -1.5460811 5.0670695 -12.774383 -0.59896344 3.648474 13.0734 -10.472006
0.499456 -0.37867 8.755862 -9.667527 5.776419 0.70775896 -2.6995454 -1.25594 -2.1210744 2.4193 -4.03013 -10.418072 4.7148232 -5.275217 -0.6542822 -3.9228692 -0.6205257
2.4754202 5.522229 -0.023137044 8.187176 0.21957615 11.4037895 -9.719555 11.94138 -7.065863 2.7874331 -0.6551325 -9.61243 -0.18044122 -0.774068 6.573521 3.217024
```

Example of embeddings for user and patient data

# Qualitative Analysis

## **Objective:**

Our aim is to observe whether we obtain communities from the “Patient-Topic Network” which consist of topics as well as users and display diversity in the topics in community.

## **Observations:**

- More than 80% of the communities consists of both users and topics.
- A lot of these communities show diversity in the topics being discussed, with regards to the forums to which they belong.

# Qualitative Analysis

Topic ID	Context
F6 T779992	Managing Side Effects Breast Cancer & treatment
F83 T773037	Not Diagnosed but Worried
F44 T758994	Breast Reconstruction
F69 T784857	Chemotherapy Before, During and After
F78 T775441	Hormonal Therapy Before, During and After

**Table 3: Cluster #01:** There are 5 topics in this cluster belonging to 5 different forums. There is no common patient (user) between these topics.

# Qualitative Analysis

Topic ID	Context
F91 T792393 + 2*	Surgery - Before and After
F96 T835504	IDC Invasive Ductal Carcinoma
F67 T796919 + 2*	Stage III Breast Cancer
F93 T784857	General Comments and Suggestions
F16 T776398	For Caregivers, Family, Friends & Supporters
F5 T793169	Just Diagnosed

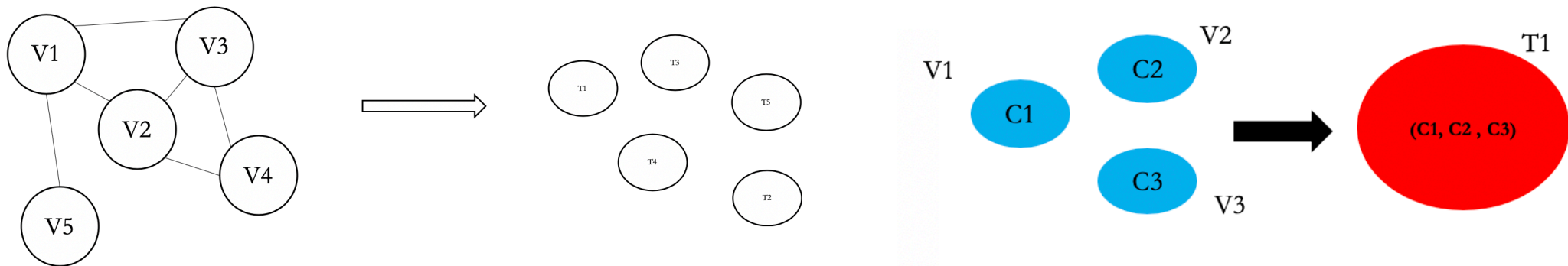
Table 4: **Cluster #02:** There are 11 topics in this cluster belonging to 6 different forums. Patient  $v^*$  from this cluster has a diverse topic interaction and tracing their journey can help the other patients in this community.



# Quantitative Analysis: Baseline Models

## Homogeneous Model (User (patient) Network)

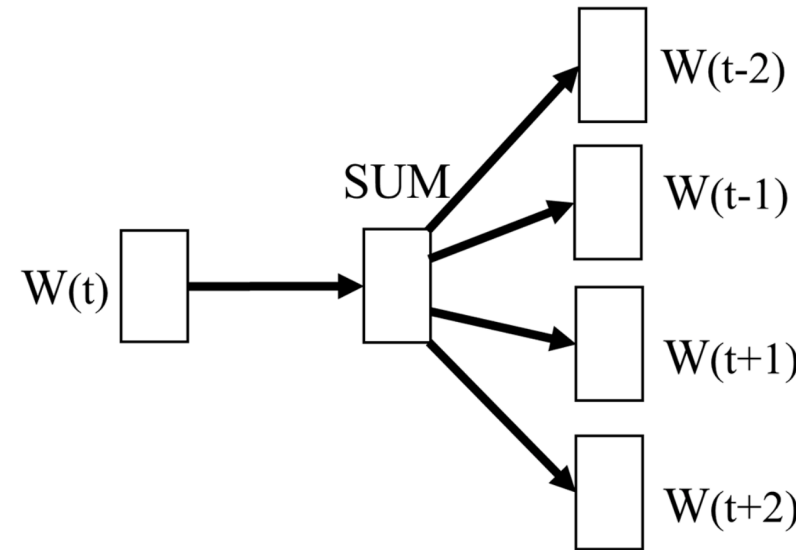
- Created a homogeneous a network graph with user data, where we draw an edge between users if they have posted in the same topic and use Node2Vec<sup>[2]</sup> to create the random walks and embedding.
- The patient feature space is mapped into the topic space by replacing the patients with the topics they had directly interacted with.
- If more than one patient has participated in the topic, the final embedding for that topic will be the centroid of all the patients' vectors who have participated in that topic.



# Quantitative Analysis: Baseline Models

- **Text Based Model (Word2Vec)**

a. Created a Word2Vec<sup>[5]</sup> model from the words extracted from the posts.



Snippet from Jang et al (2019)

# Quantitative Analysis : Evaluation

## Coherence :

- Topic Coherence is a measure that takes into consideration the degree of similarity between items in the topic and it is often used to measure the quality of the vectors in embedding methods<sup>[5]</sup>.

	Coherence(NPMI)
Patient - Topic Network	<b>0.481</b>
User (patient)Network	0.237
Word2Vec	0.294

Table 5: Comparing the performance of the patient-topic network using Normalized Pointwise Mutual Information (NPMI) to measure coherence

# Quantitative Analysis : Evaluation

## Comparison with Reference Data :

- Manually categorized posts<sup>[6]</sup> to identify actionable topic clusters from BreastCancer.org.
- We used that manual set as a reference to evaluate our method using information retrieval evaluation metrics.

	Precision	Recall
Patient - Topic Network	<b>0.643</b>	<b>0.588</b>
User (patient)Network	0.428	0.314
Word2Vec	0.507	0.422

Table 6: The performance of the different methods in identifying communities with respect to the reference dataset. Here higher precision-recall values represent better identification of the communities as defined in the reference dataset.

# Quantitative and Qualitative Analysis : Summary

- Our novel approach enables us to cluster topics on online forums which would've otherwise been tagged to be unrelated
- Using the patients as variables helps extract richer information
- We can infer both implicit and explicit relations in from the communities obtained from a heterogeneous graph
- The “patient – topic network” shows its superiority to the other two baseline models in context to this data

# Tools and Algorithm

## Tools

### Languages:

1. Python
2. SQL

### Methods:

1. Network Analysis
2. Natural Language Processing
3. Machine Learning

### Platform:

1. Jupyter Notebook
2. phpMyAdmin

## Algorithm

**LearnFeatures** (Graph  $G = (V, T, E)$ ,  $k, d, r, l, p, q$ )

$\pi = \text{PreprocessModifiedWeights}(G, p, q)$

$G' = (V, T, \pi)$

**node2vecWalk** (Graph  $G' = (V, E, \pi)$ , Start node  $u$ , Length  $l$ )

**walk** = [ $u$ ]

for walk\_iter = 1 to  $l$

$\text{curr} = \text{walk}[-1]$

$V_{\text{curr}} = \text{GetNeighbors}(\text{curr}, G')$

$s = \text{AliasSample}(V_{\text{curr}}, \pi)$

    Append  $s$  to walk

return walk

**Obtain** Embeddings

Input:  $\{e_1, e_2, e_3, \dots, e_n\}$  Set of entities i.e. embeddings in our case

$K$ : number of clusters (elbow method)

*MaxIters* : Limit of iterations

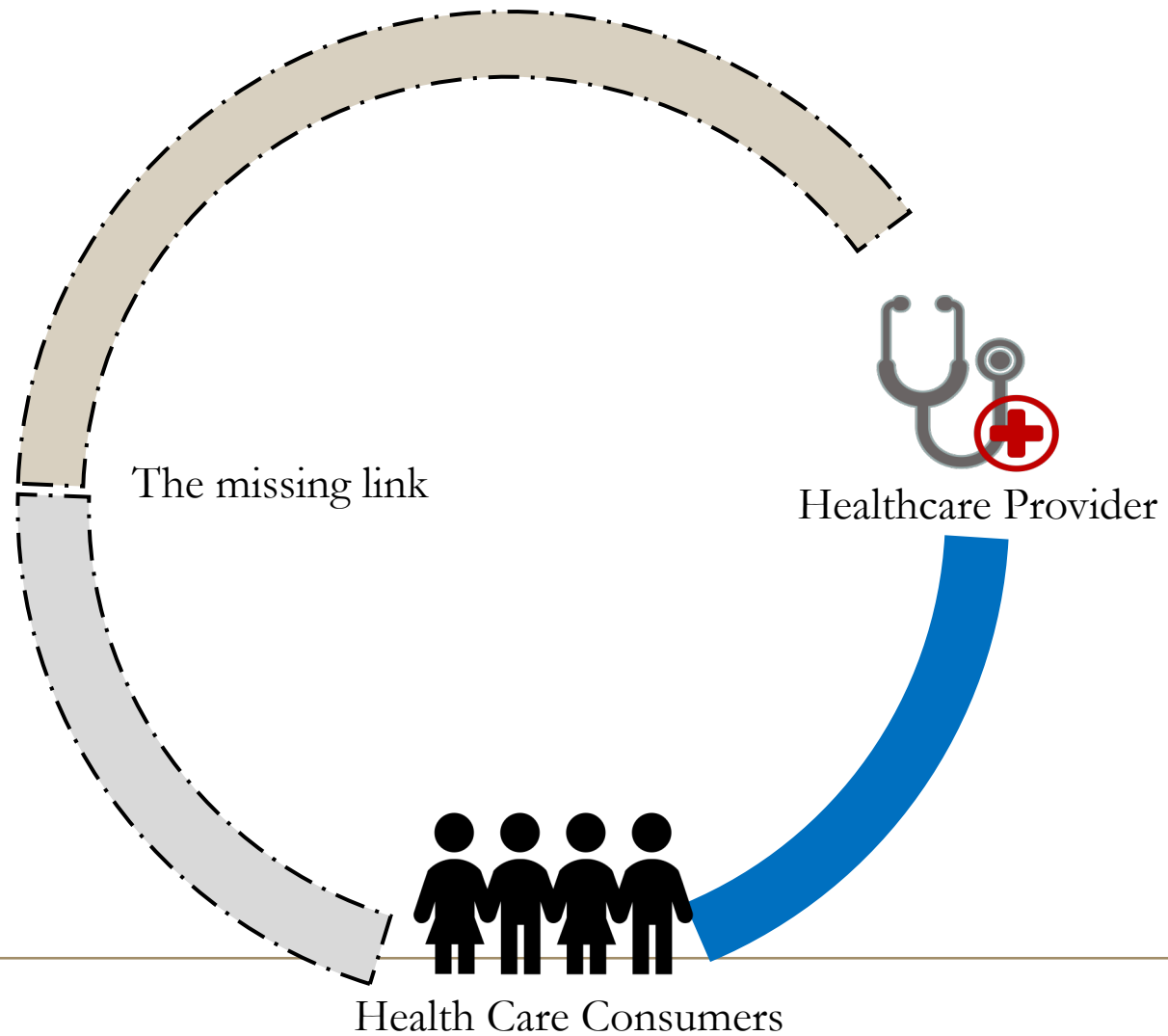
Output:  $\{c_1, c_2, c_3, \dots, c_k\}$

**Calculate** distance of nodes to centroids

**Obtain** members of each cluster

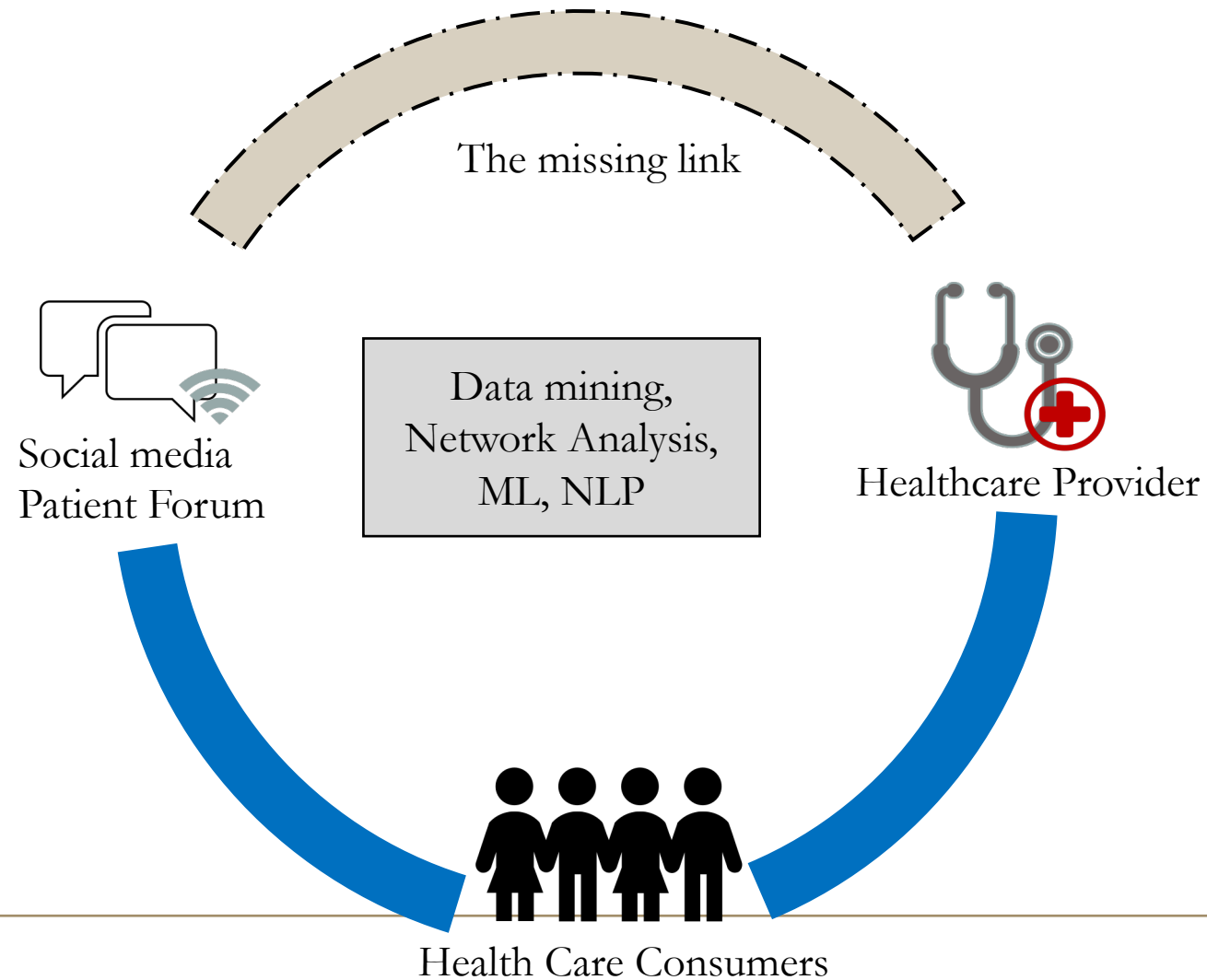
Partial snippet from Grover et al (2016)

# Contribution



Slide courtesy: Dr. Josette Jones

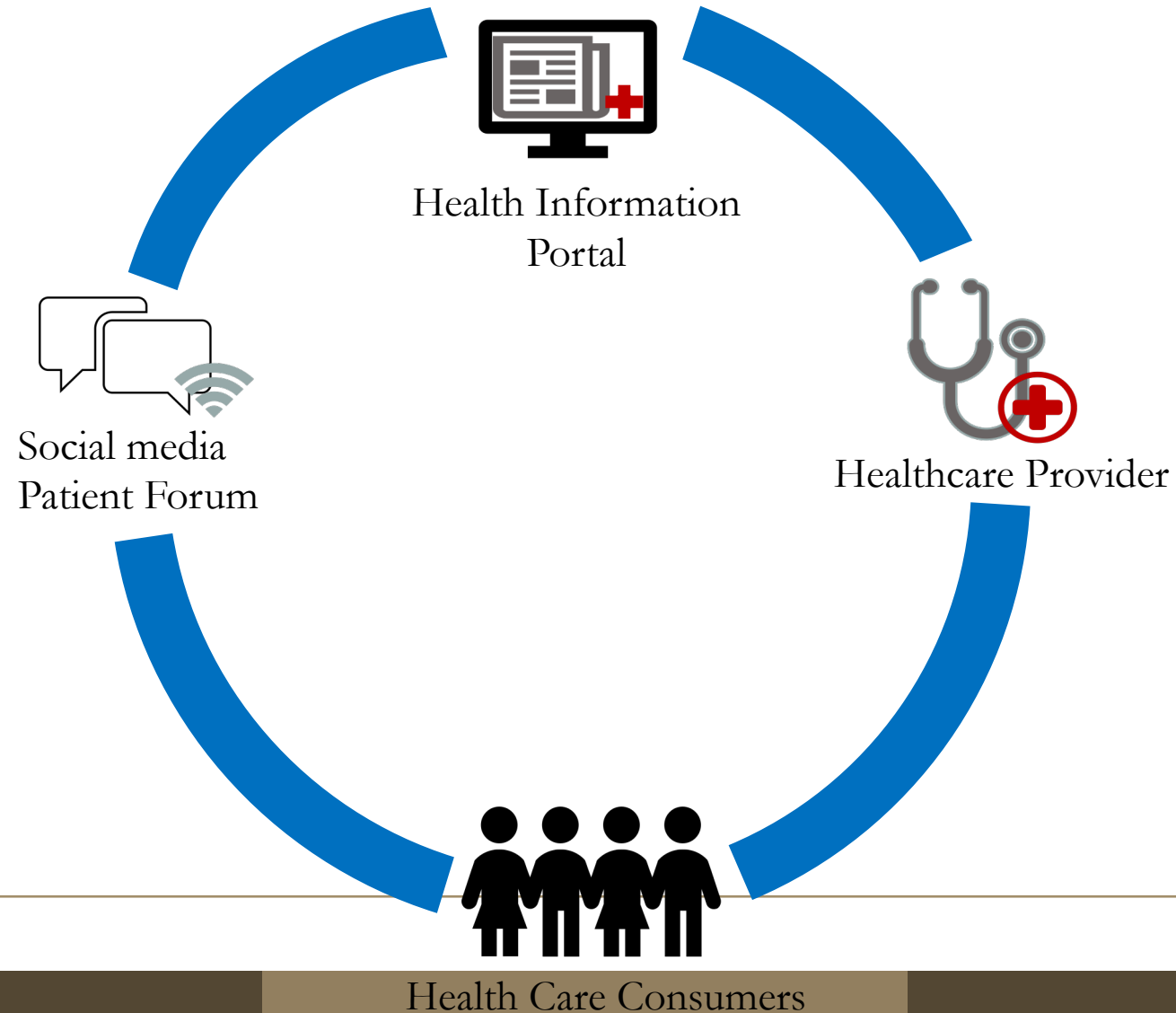
# Contribution



Slide courtesy: Dr. Josette Jones



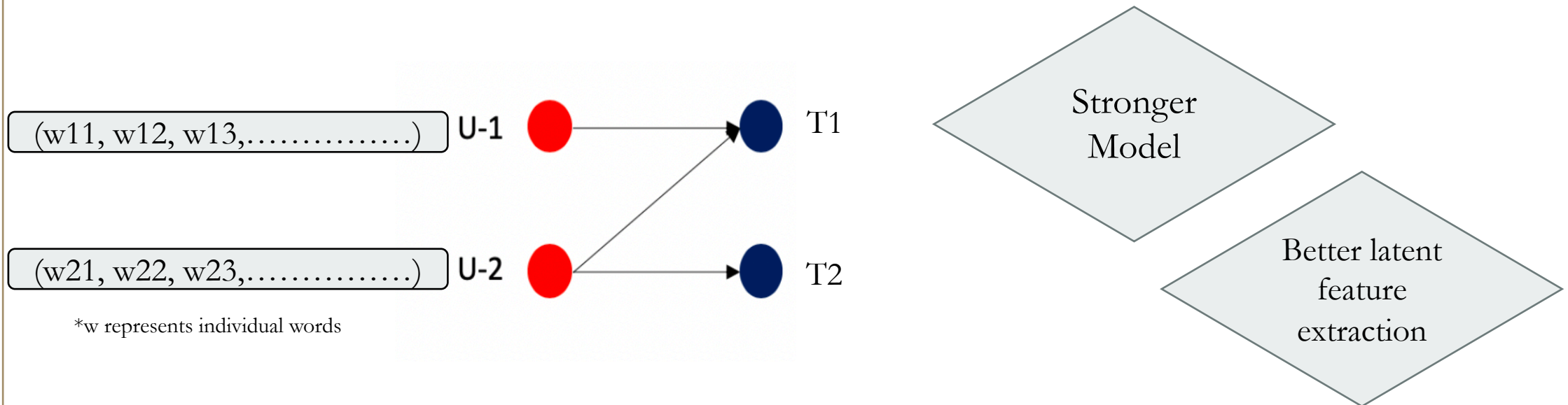
# Contribution



Slide courtesy: Dr. Josette Jones

# Future Work

- Online social media health platform: users “post” their experiences and opinions.
- The patient-topic network can be further strengthened by including the words from the posts.
- In natural language processing – machine learning models, we can establish a strong connection between two users if they have used common words.



# Future Work



Involve clinicians to evaluate the user-topic communities



Create a knowledge base of the social and personal problems related to Breast Cancer



Integrate this knowledge with physicians existing work-flow



Create a tool as a web-based platform to deploy across various medical facilities in the state of Indiana and analyze the clinical benefits of the study



Through our method we wish to create automatic generation of timeline or a road map for an individual patient based on the current diagnosis



Extend this hypothesis and research concept to other chronic diseases and their treatment work-flow

# Conclusion



This novel approach of feature representation framework allows us to connect patients and topics and get a holistic view.



We observe that the newly constructed vectors can preserve the structure of the network as well as identify new relationships connecting similar patients or patients with similar topics, even when these relationships are not explicitly depicted in the data and understand the superiority of this approach.



We evaluate our model by showing the coherence of the new relationships and better performance compared to other similar methods.

# Publications

- **Maitreyi Mokashi**, Enming Zhang, Josette Jones, Sunandan Chakraborty.  
2020. Extracting Features from Online Forums to Meet Social Needs of Breast Cancer Patients. In ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)(COMPASS '20), June 15–17, 2020, , Ecuador. ACM, New York, NY, USA, 10 pages.  
<https://doi.org/10.1145/3378393.3403652>

# References

1. BiNE: Bipartite network embedding, Ming Gao, Leihui Chen Xiangnan He, and Aoying Zhou, 2018
2. Node2Vec: Scalable feature learning for networks, Aditya Grover, and Jure Leskovec, 2016
3. A comparative analysis of community detection algorithms on artificial networks, Zhao Yang, Rene Algesheimer, and Claudio J Tessone, 2016
4. Efficient estimation of word representations in vector space, Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013
5. Text coherence new method using word2vec sentence vectors and most likely n-grams, Mohamad Abdolahi Kharazmi and Morteza Zahedi Kharazmi, 2017
6. Novel Approach to Cluster Patient-Generated Data Into Actionable Topics: Case Study of a Web-Based Breast Cancer Forum, Josette Jones, Meeta Pradhan, Masoud Hosseini, Anand Kulanthaivel, and Mahmood Hosseini, 2018
7. [Podcast] Bridging the Communication Gap Between Doctors and Patients. Retrieved July 23, 2020, from <https://healthcaresuccess.com/blog/patient-experience/bridging-the-communication-gap-between-doctors-and-patients.html>
8. Thakkar, S. (2018, October 09). Bridging the Communication Gaps Between Patients and Providers. Retrieved July 06, 2020, from <https://myvetahealth.com/bridging-communication-gaps-patients-providers/>

# References

9. Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ. 1995;152(9):1423-1433.
10. Bull SA, Hu XH, Hunkeler EM, et al. Discontinuation of use and switching of antidepressants: influence of patient-physician communication. JAMA. 2002;288(11):1403-1409. doi:10.1001/jama.288.11.1403
11. Ivedix, Author: IVEDIX Website: [@iVEDiX](http://test.ivedix.com), IVEDIX, A., & <Http://test.ivedix.com>, W. (n.d.). Mobile Healthcare Bridges Providers-Patient Gap. Retrieved July 06, 2020, from <https://ivedix.com/mobile-healthcare-bridges-providers-patient-gap/>
12. Francis, V., Korsch, B. M., & Morris, M. J. (1969). Gaps in doctor-patient communication: Patients' response to medical advice. New England Journal of Medicine, 280(10), 535-540.
13. Goudge, J., Gilson, L., Russell, S., Gumede, T., & Mills, A. (2009). Affordability, availability and acceptability barriers to health care for the chronically ill: longitudinal case studies from South Africa. BMC health services research, 9(1), 75.
14. Mira, J. J., Guilabert, M., Pérez-Jover, V., & Lorenzo, S. (2014). Barriers for an effective communication around clinical decision making: an analysis of the gaps between doctors' and patients' point of view. Health Expectations, 17(6), 826-839.

# References

15. Gu, G., Zhang, X., Zhu, X., Jian, Z., Chen, K., Wen, D., ... & Lei, J. (2019). Development of a consumer health vocabulary by mining health forum texts based on word embedding: semiautomatic approach. *JMIR medical informatics*, 7(2), e12704.



# Acknowledgements

- Sunandan Chakraborty, Assistant Professor, Human-Centered Computing Dept., SOIC, IUPUI.
- Josette Jones, Director, BioHealth Informatics Dept., SOIC, IUPUI
- Jiaping Zheng, Assistant Professor, BioHealth Informatics Dept., SOIC, IUPUI
- Enming Zhang, PhD Candidate, BioHealth Informatics Dept., SOIC, IUPUI
- Dr. Jones's and Dr. Chakraborty's research groups
- Elizabeth Cassell, HCC Dept and The School of Informatics and Computing, IUPUI

Thank you

Questions?

Maitreyi Mokashi  
MS Applied Data Science,  
School of Informatics and Computing, IUPUI  
[mmokashi@iu.edu](mailto:mmokashi@iu.edu)

